Journal of Experimental Education

Volume IX | March, 1941 | Number 3

A STUDY OF HIGH SCHOOL AND COLLEGE SUBJECT MATTER PATTERNS FOR FOUR GROUPS OF UNIVERSITY STUDENTS

William N. Atkinson

Previous Studies

Through the first decade of the present century there was relatively little questioning of the traditional classical requirements for college entrance. In 1911, the Committee of Nine on the Articulation of High Schools and Colleges (6) recommended the requirements for college entrance which have since been more or less standard. These requirements consisted of nine specified units, three in English, two in a single foreign language, two in mathematics, one in social science including history, and one in natural science. Two additional units were to consist of advanced work in academic subjects to complete the requirement of a second major of three units. Four units were left as a margin to meet individual needs.

In 1919, Colebank (5: p. 149) stated the conviction which was growing among more liberal school men that

> "No convincing evidence has made certain that a four year secondary school course, consisting largely of quite formal studies in foreign language and mathematics, taught by the methods that have long been customary in preparatory and other secondary schools, actually lays a better foundation for successful college work and for subsequent success in life than courses of a different character. Many colleges insist on a certain number of units of Latin or other foreign language. They should be expected to present evidence of a convincing nature that the enforcing of this requirement enables them to secure students with best preparation for college work, such results being due to the educational value of the subject."

About this time a new basis for the selection of students appeared in the form of psychological and other aptitude tests. The attention of investigators was directed toward a number of factors possibly predictive of college success. Those chiefly studied were intelligence as measured by some test or battery of tests, high school scholastic average or rank in class, and the amount of credit in specified high school fields.

As might be expected, there was not entire unanimity in the findings regarding any factor. However, there appeared to be less evidence in support of the predictive value inherent in the amount of high school credit in any field than in support of the predictive value of either intelligence or high school rank. After a study of the scholastic records of 450 freshmen in Syracuse University, May (14: pp. 436, 437) concluded that

> "By knowing the intelligence of a student and knowing the time he spends in study we can predict his honor points with but slightly greater error than we could if we knew also the units he offered and the average grade he made in them . . . Aside from statistical treatment there are other considerations that lead us to believe that the amount of preparatory school or high school work is a very poor index of college success."

Gowen and Gooch (10), in a study of 927 graduates of the University of Maine, found that, compared on either the amount or the quality of the work done in each, high school subjects were about equal to each other in predictive or training value for the freshman year of college. Odell (16) made an extensive study of about 2000 freshmen in a number of colleges and found some predictive value with respect to college grades in the number of credits in specified high school subject fields. However, in almost every case, the general high school average, the I.Q., or

both taken together, were superior to specific subject patterns in predictive value.

In 1923 Gebhardt (9) found no significant evidence, in a study of students at the Colorado State Teachers' College, to indicate that any one subject or group of subjects in high school is of greater value than another as preparation for college work. In 1924, MacPhail (13) selected twenty-seven pairs of students of equal rating on psychological tests, of which one member had taken four years of high school Latin and the other none. The average freshman grades favored the Latin members of the pairs but the difference was not statistically significant.

Bolenbaugh and Proctor (1) found little difference between the academic achievements of two groups of Stanford University men, one of which had three or more units of vocational subjects in high school and the other no more than one. Brammell (2) made a similar analysis of the records of 163 freshmen entering the University of Washington in 1925 with five or more vocational units and of 353 with no more than one non-academic credit. The first year average favored the academic group slightly but the four year averages were identical.

Sorenson (19), in a study of the success of Northwestern University freshmen, concluded that three or four years of high school Latin were definitely predictive of college success. French and mathematics appeared to differentiate students less efficiently. Kent and Schreurs (12) studied the relation of mental alertness scores, rank by quarters in high school class, amount of high school English and mathematics, and general first year college grades to success in college English and mathematics courses for 228 men and 296 women at Northwestern University. They found the amount of high school English to have a slight negative value for predicting success in college composition. The amount of high school mathematics had positive value for predicting success in trigonometry but not in algebra.

Brammell (3), in 1932, reporting on the findings of the National Survey of Secondary Education in regard to the entrance requirements of 517 colleges and universities, listed the average prescription at about ten units. In the period from 1924 to the time of the Survey, 209 requirements in the academic fields had been decreased or eliminated. On the other hand, 165 had been increased. Many provisions for regular admission other than by high school transcript were listed, but that remained the sole method in 80 per cent of western colleges and the chief method in most sections.

The study of the significance of various elements of the high school pattern for college success has not been discontinued. Douglass (8) summarized studies up to 1932 and reported on his own study at the University of Oregon (7) of 387 graduates of the class of 1930. He found that (8: p. 143) "Not only was there little or no relation between pattern of high school credits and college success, but apparently there is little relation between the mastery of any particular subject and college success." He concluded (8: p. 147) that "The author of this study feels certain that, a few decades hence, our present practice of selecting college entrants on the basis of minimum credits in certain fields of high school credits will seem a curious and inexplicable anomaly." Five years later, however, he joined with Smith (18: p. 695) in reporting their study at the University of Minnesota to say ". . . if there is any subject in high school the pursuance of which tends more than any other study to better marks in college, that field is Latin."

Whelden's intensive study of five successive classes at Yale indicated that greater quantities of Latin taken in private schools or after entering Yale College led to better scholarship in subjects in which there was identity or great similarity of elements. But (20: p. 497) "There was no evidence at all of any value of a more extended study of Latin as an intellectual discipline, serving to extend the scope of intellectual capacity in whatever field it might be applied." On the other hand, Ross (17) found that among 590 students in three classes at Allegheny College, those having more high school language did better college work.

Nelson (15) discovered that among 2000 students at Iowa State Teachers' College foreign language was the only field in which the amount of credit earned in high school showed a significant relationship to college achievement. Clark (4), in a study of 1431 women and 1182 men at Northwestern University, found Latin to be the only high school subject having any predictive value for college success.

In a study of first year achievement of 861 members of the entering class of 1934 at the State University of Iowa, Hammer (11) found that college success is more dependent upon the qualities of the student than upon the high school subject pattern. Accumulation of credit in a specific subject field had little relation to college success in that field; and general success which may be associated with foreign language and mathematics was more easily explained in other ways than by assuming the inherent value of the subject as a determiner of college success.

Summary of Previous Studies

Investigations of the importance of the high school pattern of subjects for academic choices and success may be classified as follows:

1. *Studies relating credit in specific high school subjects to general college average for a given number of semesters of college work, or to success in certain college fields, or to both.*

After due allowances have been made for ability differences, few of the relationships discovered prove significant, although some special claims are made for Latin, French, and mathematics. Most of the investigators finding such relationships are cautious in asserting that they are causal. Even in pairing studies it is questionable if all important factors other than those under consideration are perfectly equated.

2. *Studies of the general college standing of students offering classical and vocational patterns.*

Such studies reveal few differences and those small, but those which appear to exist favor the classical groups. However, the negative evidence from a number of extensive and carefully conducted studies in which no such differences were found cannot be overlooked.

3. *Studies of the relation of high school and previous college patterns to success in certain professional colleges or divisions.*

Very little appears to have been done in this field. There is some evidence that certain high school courses contribute to success in engineering but that the pattern is unimportant for law.

4. *Studies of the general intellectual or academic quality of students who express their greatest interest in, or actually enter certain specialized courses.*

The results of such studies are contradictory. Contradictions may be due to the unreliability of reported interests, to differences in admission requirements for various professional curricula in institutions where studies were made, or to differences in techniques employed in the studies. Present evidence does not appear to establish real differences among students preparing for professions for which university training is required.

Purpose of This Study

The purpose of the present study was to discover the characteristics of the pattern and quality of educational experiences which might identify each of several groups of university students. Those selected for the study were those majoring in English, law, medicine, and engineering. The investigation naturally divided into two parts, indicated by the questions: 1. What are the characteristics of students of each of these groups in relation to their relative success in the curriculum chosen? 2. How do the characteristics of students in one group differ from those of students in the others?

Collection of Data

Data for the study were secured from several types of permanent student records at the State University of Iowa. Most of these were located in the registrar's office. The groups chosen were selected as representative of curricula which differ considerably and which enroll a number of students sufficient to provide enough cases without covering too great a span of years. The greatest number of classes represented in any group was thirteen.

In this study, students were classified as successful if they completed the work of their chosen curriculum and received the appropriate degree. Students were classified as unsuccessful if they dropped the course before beginning the third year, with grades averaging less than C, or 2.0. This definition was used because total grade points equivalent to a C average are required for graduation from most colleges of the University. Unsuccessful students were placed, for purposes of analysis, in two supplementary groups, engineering and liberal arts.

Composition of the Groups

The total number of students included in this investigation was 1623. Of these, 289 completed the four year course in liberal arts

with a major in English; 251 completed the course in law; 317, the medical course, and 323 received degrees in engineering. The remainder included 176 students who were unsuccessful in engineering and 267 who were unsuccessful in the liberal arts course.

There were no women in either the successful or the unsuccessful engineering group. Women numbered only two per cent of the law and medical groups. Twenty-five per cent of the unsuccessful liberal arts group and 71 per cent of the English group were women. A relatively larger percentage of women than of men persisted in college attendance beyond the second year, since only one fourth of the unsuccessful liberal arts group consisted of women, as compared to slightly more than one third of the entire group entering at the same time.

Age at High School Graduation

The mean age of successful men students at high school graduation was about 17.7 years in each group. In all groups it was found that students who ranked high in their high school classes were, on the average, younger than inferior students at high school graduation. The mean age of women at high school graduation was below that of men in the same group. There were no students of either sex in the entire study who were graduated from high school below the age of fifteen or above the age of twenty-nine.

Size of High School Attended

Size of high school attended appeared to have little influence upon choice among the university curricula included in this study. The school size distributions for the failing groups differed so little from the others that one may also conclude, in general, that the size of high school attended has little effect upon college success or failure. Sub-groups composed of students from large schools (over 600 students) showed an advantage over sub-groups from smaller schools in qualifying examination percentile ranks and in first year grades in college. This was true for all major groups except English, in which the large-school group ranked above the small-school group (under 100 students) but below the group from schools of medium size. However, these differences tended to disappear during the course, and law was the only field in which the large-school group showed an advantage in final grades. It seems reasonable to conclude from this, that although schools in general do differ by size in the preparation they give their students for writing entrance examinations and doing first year college work, such advantages or handicaps do not persist throughout the college course.

Rank in High School Classes

The law, medicine, and engineering groups differed little in the percentage of members ranking in each third of their high school classes. The ratio of students of these groups sectioned by rank in high school class was approximately 60:30:10, from the highest third to the lowest. The English group averaged a higher rank in high school, with a ratio of 78:20:2. The unsuccessful groups resembled each other in this respect, with ratios approximating 20:50:30. From these ratios, it is evident that there is a relationship between class rank in high school and chances of success in college.

It may be deduced from the ratios given that rank in graduating class does influence the decision of students in regard to college entrance. Even in the unsuccessful groups of college students, fewer than one third of the number had ranked in the lowest thirds of their respective high school graduating classes. It is also evident that class rank by thirds offers no positive assurance of either success or failure in any of the areas studied, since low-third students were found in each successful group and high-third students in both unsuccessful groups. The latter circumstance may be explained by the fact that the definition of unsuccessful students makes no allowance for students of satisfactory ability who are unable to remain in college or to do work of adequate quality because of ill health or other factors which cannot be determined.

Delay in College Entrance

The smallest percentage of students delaying college entrance after high school graduation was in the law group, and the highest in the engineering groups. There appeared to be no particular reason for the difference. It is of some interest to note that about one fourth of all students who enter the University were out of school a semester or longer after high school graduation. However, little evidence appeared in this study to indicate that this delay affected college success in any way.

Qualifying Examinations

The mean qualifying examination percentile ranks for the law, medicine, and successful engineering groups were 63, 62, and 64, respectively. This is as near equality as could be expected. Variances were also equal. The mean rank of the English group was 71, significantly higher than those of the other groups. That of the unsuccessful engineering group was 42; and of the unsuccessful liberal arts group, 31. The differences in the means of successful and unsuccessful groups indicate that the qualifying examination has some value as an index of college success or failure. However, the fact that every successful group included students ranking in the lowest decile on the qualifying examination shows that it is far from absolute as a predictive instrument.

Relative Scholastic Ability

The differences in the ratios of students ranking in the highest thirds of their high school classes, in mean qualifying examination percentile ranks, and in mean one year and two year grade point averages, all point to the English group as scholastically superior to the other successful groups included in this study. There is evidence that this superiority is not a result of the preponderance of women in the English group in the fact that the men of the English group averaged higher on each of these measures than the women. The data which support the conclusion as to the superiority of the English group point just as definitely to the scholastic equality of the law, medicine, and engineering groups.

College Grades in Required Subjects

With few exceptions, the grade means of all sub-groups in required courses in the first and second years of college varied from the grade means of these groups in all subjects taken at the same time in the direction of a 2.0, or C average. This leads one to question whether apparent differences in the abilities of student groups in required subjects, as contrasted with their abilities in all subjects considered together, may not be misleading. C grades may be more common in these courses, where large classes and routine instruction make reliable measures of individual accomplishment difficult to obtain. A preponderance of C grades would have the effect of undergrading the superior students and overgrading the inferior, which would explain the conditions noted. Achievement of outstanding grades would represent unusual promise in the field and might be extremely valuable in the selection of students in appropriate majors, especially English.

Correlations

Qualifying examination percentile rank was correlated with final grade averages for English majors more highly than with grades of engineers, .55 and .43, respectively. Correlations with grade averages in law and medicine were smaller, .35 and .33. First year grades showed correlations of .86 and .80 with total English and Engineering grades, but a spurious element is involved in the inclusion of first year grades in the total from which final averages were calculated. Correlations of first year grades with final grade averages in law and medicine, involving no spurious element, were .49 and .47. Correlations of two year pre-professional grades with final grade averages were .50 for law and .53 for medicine.

Size of high school attended showed no more than chance correlation with final grade averages in any field. Rank in high school class, however, showed significant correlations with final grade averages in every successful group. Values ranged from .40 in English to .25 for law, with that for engineering .26, and for medicine, .31. The latter was almost as high as the corresponding correlation of qualifying rank with final grade averages.

High School Subject Matter Patterns

Mean amounts of high school English ranged from 8.01 semesters for the English group to 7.66 for the engineering groups. The greatest difference was less than half a semester. In language credit the English group also ranked first, with a mean of 5.12 semesters. The law and medical groups earned means of 4.89 and 4.82 semesters, respectively; the successful engineering group, 3.63; the unsuccessful liberal arts group, 2.96; and the unsuccessful engineering group 2.83 semesters. The maximum difference in this case was two and one fourth semesters.

Mean credit in social studies ranged from 6.25 semesters for the law group to 5.64 for the successful engineering group. Mean mathematics credit covered a range of a semester, from 5.85 semesters for successful engineering

students to 4.82 for English students. The engineering group was high and the law group low in mean amounts of credit in science, with 5.16 and 4.28 semesters, respectively. Credit in the non-academic field ranged from a mean of 5.14 semesters for the unsuccessful liberal arts group to 2.96 for the medical group, a difference of more than two semesters. The average total of high school credits ranged from 31.90 semesters for the law group to 31.08 for the unsuccessful engineering group.

When the unsuccessful liberal arts group is compared with the three successful groups which began their courses in the College of Liberal Arts, it is seen that the mean subject matter pattern of the former group differs notably from the others only in the smaller amount of credit in language and the greater amount of credit in non-academic subjects earned in high school by the unsuccessful group. These differences were about two semesters of credit each. None of the differences in mean amounts of credit in any field among the three successful groups was as great as one semester. If the engineering groups are added to the comparison, the greatest difference in mean amounts of language credit is extended to two and one fourth semesters and that in mathematics to more than a semester. Credit in non-academic subjects is still within the two semester range.

Among all the groups considered, high school credit varied greatly only in the language, the mathematics, and the non-academic fields. The English, law, and medical groups earned the highest mean amounts of credit in language. The unsuccessful groups were characterized by greater amounts of non-academic credit. The engineering groups averaged more mathematics credit, and the law group a small amount more of social studies credit. The engineering and medical groups took, on the average, more high school science than the other groups.

While these differences indicate that possibly some preparation was made in high school for the curricula selected in college, both individual and sub-group variations in each successful group equalled or exceeded the differences in the means of the major groups. In some cases, the pattern of preparation for superior students within a group differed from that of the lower ranking students of the same group in the same manner as that of the entire group differed from those of other groups. This was not universally true, however, for the high ranking students of the medical group averaged less credit in high school science than the low ranking medical students. While high ranking students in English earned more credit in foreign language than those ranking low, both sub-groups averaged less than the English group as a whole.

From the data presented, it appears that there is no pattern of high school subject matter that prepares students for success in one of the fields covered by this study more definitely than for success in any other. The sole possible exception to this is in the case of additional mathematics and science for engineering students. On the other hand, it appears that greater amounts of language and mathematics credit, and smaller amounts of science and non-academic credit distinguish, to some extent, the better from the poorer students in any field. Again, science for the engineering group must be excepted. Amounts of credit in high school English and social studies appear of small significance, possibly because there was little variation in credit in these fields.

Relatively large curricular offerings in language and mathematics, coupled with reduced college entrance and high school graduation requirements in these areas, and small requirements, along with increased offerings, in non-academic subjects, account for the possibility of greater variation in the amounts of credit in these fields offered for college entrance. Since greater variation does occur in those fields in which it is possible, it seems reasonable that greater weight should be given to the small variations noted in other fields.

On this basis, it appears that students entering engineering have, on the whole, made some special preparation for this curriculum by taking additional mathematics and science in high school. Students who are unsuccessful in engineering have, on the whole, less preparation in mathematics than those who are successful. Law students, on the other hand, have tended to take more language and social studies than the average for all students; medical students more language and science; and English students, more English and language. Even in the successful groups, there is some evidence that the better students show greater differences in this respect than the

poorer. It should be kept in mind, however, that these differences, while consistent, are probably too small to have statistical significance, and might be accounted for by chance.

Variations in High School Scholarship

An additional matter of interest in connection with the high school subject matter pattern is the number of students ranking higher or lower in each high school field than in all high school subjects taken together. The greatest number of high ranking students in the English major group was in the English field in high school; in law, social studies; in the successful engineering group, mathematics; and in both unsuccessful groups, nonacademic. The greatest number of low high school grades for the English major group was in mathematics, with the greatest net number in science. In law, the most low high school grades were in mathematics; in medicine, language; in both engineering groups, English; and in the unsuccessful liberal arts group, mathematics, with the greatest net number in language. These results may show a trend toward better or poorer work in certain high school fields for members of different university groups, but in every successful group each high school field was represented by both high and low grades.

General Conclusions From This Study

The general conclusions drawn from this study may be stated as follows:

1. *The existence of small variations in the amounts of credit earned in high school fields in which requirements are high, and greater variations in fields with more liberal requirements, indicates the possible presence of some relationship between the curriculum pursued in the university and the high school subject matter pattern.*

While the specific relationships between amounts of credit in various high school fields and the choice of college curriculum are small, they are consistent among the four curricula included in this study. Since the possibilities of variation in the high school pattern are restricted, any variation for a considerable group of students suggests the possibility of the operation of some selective factor. It must not be assumed, however, that any relationships which appeared were absolute or essential, as differences of this size might well occur by chance.

2. *The relative number of students receiving higher or lower grades in certain high school fields than their average in all, points to the possibility of special relationships between certain high schools and college fields.*

The piling up of frequencies of higher grades in certain high school fields and of lower grades in others to show a trend in the distribution of special interest or ability for a college group appears worthy of notice and may justify further study. However, the presence of many high and low grades in other high school fields necessitates extreme caution in advancing any generalizations on the subject on the basis of this study.

3. *The general scholastic ability of the student, as measured by rank in thirds of high school graduating class, rank on college entrance examinations, and grades in preliminary college courses, appears to be more important for success in any of the curricula included in this study than specific characteristics of the high school subject matter pattern.*

Every group included in the study offered satisfactory evidence for this conclusion. It must not be implied, however, that this generalization could be applied to predict success or failure in an individual case.

Subsidiary Conclusions

In addition to the general conclusions listed above, several subsidiary conclusions appear tenable:

1. Successful English students, as a group, are superior in general scholastic ability to groups of successful students in law, medicine, and engineering.
2. There are no real differences in general scholastic ability among groups of successful students in law, medicine, and engineering.
3. Unsuccessful students, as groups, are inferior in all measures of scholastic ability used in this study to corresponding groups of successful students.
4. The overlapping of measures of scholastic ability for individuals in successful and unsuccessful groups, and for individuals ranking high and low scholastically within successful groups, indicates that indices of achievement in any of the curricula included in this study are extremely unreliable when applied to individuals.

5. There was some evidence in this study that the size of high school attended may affect relative rank on qualifying examinations or in first year grades. However, any advantage resulting from graduation from a larger high school is reduced considerably or even eliminated by the end of the university program.
6. Rank by thirds in high school graduating class, qualifying examination percentile rank, and grades earned in the first and second year of college show, for all groups, correlations of an approximate range from .25 to .50 with averages of all grades in the curriculum.
7. Since students ranking in the lowest thirds of their high school classes did not appear in either successful or unsuccessful groups in as large proportion as one third, it may be concluded that relatively fewer of these students than of students ranking higher in their high school classes enter college.
8. The proportion of women in the unsuccessful liberal arts group was smaller than in the entire entering group, indicating a relatively lower college mortality for women than for men.
9. The presence, in all successful groups, of considerable numbers of students with educational characteristics quite similar to those of failing groups suggests that some quality, which might be called persistence, not revealed by any of the educational measures recorded in this study, may be an important factor in college success.

General Implications of the Study

From this study and others in the area which have been reported, it appears reasonably safe to point to a few generalizations concerning the relationship of high school and college subject matter patterns as probabilities, even though complete proof may be lacking.

In the first place, students who take more high school language and mathematics are better students, as a group, at least in the first years of college, than those who take less. This may be explained in four ways: 1. A difference in the subject matter which gives inherently better background for any scholastic task to the first group; 2. Differences in the learning requirements and in the academic competition encountered which develop better study habits in the first group; 3. A selective factor, operating at the high school level, to give students of greater natural ability the first type of preparation; and 4. The weighting of certain measures of ability and achievement at the college level with matters not properly related to the immediate scholastic task, to favor the first type of preparation.

Very likely each of these conditions is operative on some occasions and to some extent. The writer believes that the operation of guidance methods at the high school level tends to place better students in academic and poorer students in non-academic curricula or subjects. Since this is true, scholastic competition is keener, on the average, in elective academic classes; teachers demand a better quality of work; and students receive more stimulation to their best efforts. It is also notable that in the academic subjects mental activity of a type more closely resembling that required in college is emphasized, while in many non-academic subjects manual skill is given the greater emphasis. Aside from this, there appears to be no reason to question Whelden's (20) conclusion that identical or very similar elements constitute the only gain for college work in the study of such a subject as Latin.

Brammell's (2) study, as well as others, shows that there is an original advantage in college for students with academic preparation. Four year grade averages, however, show academic and non-academic groups of equal mental ability to be practically identical in scholarship. The extent to which the first year grades in college are truly representative of the ability and achievement of the two groups may thus be questioned. In addition to other factors, there may be an actual bias, unrelated to academic merit, in favor of the academic preparation, operating more strongly in early than in later years of the college course. Some of the pairing studies seem to contradict this, and the statement must be qualified according to one's estimate of the reliability of such studies.

Another question of importance in this area is that of specialized abilities or aptitudes. These do exist, of course, but it is quite possible that they are emphasized beyond the point justified by their true importance. From this study, it appears that the majority of

students in law, medicine, or engineering could have entered any of the three curricula with practically the same chances of success as in the one they did enter as far as their ability to meet the requirements of the course is concerned. Interest, of course, is another matter. It must also be noted that these statements are made in relation to success in professional schooling, not professional practice, which may differ. However, it seems safe to encourage a student with a strong professional interest to follow it, provided he is capable of achieving success in any profession at the same level.

One of the most impressive findings of the present study, in the opinion of the writer, is the success, in each group, of students who, by every measured index, should have failed. One may still question whether they should have been permitted membership in their professions, and what their ultimate success may be. However, it is evident that these students possessed some combination of personal qualities which enabled them to survive academic rebuffs which would have discouraged most students. The writer believes, on this evidence, that one should be very cautious in stating dogmatically that any given professional goal is impossible of achievement for any student, although it may be seen as improbable.

The findings of research definitely are not such as to make it inadvisable to encourage a high school student to follow a major interest in a non-academic field, even though he may later enter college. It does seem advisable, however, that he should be required to exert strong and sustained mental effort in the work that he does pursue and that he should not be shielded from the results of his scholastic limitations. A realistic view of his mental abilities, and training in meeting defeat and persisting in spite of it may well be of greater value to a student of mediocre ability than any particular pattern of subject matter he may study.

Bibliography

1. Bolenbaugh, L., and Proctor, W. M. "Relation of Subjects Taken in High School to College Success," *Journal of Educational Research* XV (Feb., 1927), pp. 87–92.
2. Brammell, P. R. *A Study of Entrance Requirements in the University of Washington,* University of Washington, Dissertation, 1930.
3. Brammell, P. R. "Practices and Problems in Improving the Articulation of High School and College," *Proceedings of the Sixteenth Annual Meeting of the Department of Secondary-School Principals,* Bulletin of the Department of Secondary-School Principals, No. 40 (March, 1932), pp. 170–196.
4. Clark, E. L. "Amount of High School Latin as an Indicator of Success in College Work," *School and Society,* XXXV (Feb. 6, 1932), pp. 189–190.
5. Colebank, G. H. "Rational College Entrance Requirements," *Journal of Education,* LXXXIX (Feb. 6, 1919), pp. 149–150.
6. Committee of Nine. "Report on the Articulation of High School and College," *Proceedings of the National Education Association,* 1911, pp. 559–566.
7. Douglass, Harl R. *The Relation of High School Preparation and Certain Other Factors to Academic Success at the University of Oregon,* University of Oregon Publications, Education Series III, No. 1 (September, 1931).
8. Douglass, Harl R. "Selecting Good College Risks," *School and Society,* XXXV (Jan. 30, 1932), pp. 140–147.
9. Gebhardt, G. L. *The Relative Value of College Entrance Subjects,* Colorado State Teachers' College, Thesis, 1923.
10. Gowen, J. W., and Gooch, Marjorie. "The Mental Attainments of College Students in Relation to Previous Training," *Journal of Educational Psychology,* XVI (November, 1925), pp. 547–568.
11. Hammer, M. L. *The Relation of High School Subject Matter Combinations to College Success,* State University of Iowa, Thesis, 1936.
12. Kent, R. A., and Schreurs, Esther. "Predictive Value of Four Specified Factors for Freshman English and Mathematics," *School and Society,* XXVII (February 25, 1928), pp. 242–246.
13. MacPhail, A. H. "Freshman Academic Achievement in College Students Presenting Four Years of Latin and No Latin," *School and Society,* XIX (March 1, 1924), pp. 261–262.

14. May, Mark A. "Predicting Academic Success," *Journal of Educational Psychology*, XIV (Oct., 1923), pp. 429–440.
15. Nelson, N. J. "A Study of the Value of Entrance Requirements at Iowa State Teachers' College," *School and Society*, XXXVII (February 25, 1933), pp. 262–264.
16. Odell, C. W. *Predicting the Success of College Freshmen*, University of Illinois Bulletin, XXXV, No. 2 (September 13, 1927).
17. Ross, C. F. "A Method of Forecasting College Success," *School and Society*, XXXIV (July 4, 1931), pp. 20–22.
18. Smith, Mary E., and Douglass, Harl R. "The Relation of High School Latin to Marks in the First Year of Arts College," *School Review*, XLV (November, 1937), pp. 695–701.
19. Sorenson, Herbert. "High School Subjects as Conditioners of College Success," *Journal of Educational Research*, XIX (April, 1929), pp. 237–254.
20. Whelden, C. H. "Training in Latin and the Quality of Other Academic Work," *Journal of Educational Psychology*, XXIV (October, 1933), pp. 481–497.

THE STUDY HABITS OF FAILING AND SUCCESSFUL STUDENTS IN THE FIRST TWO YEARS OF COLLEGE

C. W. BROWN
University of California

INTRODUCTION

An accurate determination of the contribution of correct habits of study to college success must await the development of quantitative methods for measuring the various study skills utilized in college. In the absence of a better method, the questionnaire technique was used in the present investigation in an effort to determine roughly the amount of variation in study methods to be found among failing and successful students. The results to be presented are merely suggestive but they assist in getting a more complete description of the factors affecting the student's ability to do work at the college level.

CONSTRUCTION AND ADMINISTRATION OF THE QUESTIONNAIRE

A questionnaire was constructed with items covering such topics as the following: (1) reading, (2) reviewing, (3) concentrating, (4) critical thinking, (5) note taking, (6) preparing for examinations, and (7) promptness in completing assignments. After two revisions the questionnaire contained 64 items.

The questionnaire was filled out by 211 sophomore students on probationary status in the College of Letters and Science[1] and by 120 students in an elementary psychology course in the same college. The two groups were comparable in the number of semesters spent in the University and in the number of credit hours of work attempted. Ample time was allowed for everyone to answer all items.

METHOD OF EVALUATING THE RESPONSES TO THE QUESTIONNAIRE

To make certain that there was no overlap between the two groups of students relative to their college scholarship, the grade point average for each individual was obtained from the registrar of the University. Arbitrary limits of —4 grade points[2] and +4 grade points were set up, and the questionnaires of all students whose college average fell within these limits were discarded. This step, although reducing the number of available cases, resulted in two groups with a decided difference in their average college achievement.

In order to find out if differences existed in the methods of study of the two groups the percentage in each group giving the same answer to each item was computed. The differences between these percentages, and the reliability of these differences, were determined and used as a basis for evaluating the study methods.

Most of the percentages to be considered in the following discussion are for what may be considered the psychologically correct answer. By "psychologically correct answer" is merely meant the answer which was given by the majority of graduate students in psychology who were asked to fill out the questionnaire. For many of the items there is no "psychologically correct answer". Disagreement concerning which of the possible answers to an item is correct was expected, but such disagreement does not invalidate the findings to be reported.

It is not practicable to include a summary of the responses to all of the items in the present report. Instead a comparison of the failing and successful students will be made in terms of the larger aspects of study methods already mentioned. Some of the items of the questionnaire were related to more than one of these aspects, and therefore may be included in more than one of the topics in the following discussion.

READING HABITS

In the table are found the data necessary for the evaluation of the different phases of study habits. Columns three and four contain the average percentage of each group giving the correct answers to all of the items in a given phase. In columns five, six and seven the critical ratios of the differences for indi-

[1] University of California, Berkeley.

[2] Grade points are assigned for each unit of credit as follows: A 3, B 2, C 1, D 0, E and F —1.

vidual items are distributed according to size and separated according to the group favored by the difference.

Twenty items of the questionnaire dealt with methods of reading. It can be readily seen that there is little difference between the answers of the two groups. In a few of the items one or the other of the groups gave a significantly greater percentage of correct answers, but on the whole the differences were small and favored each group about the same number of times. The data do not merit placing much confidence in any individual item. However, the consistency of the small differences in so many items supports the conclusion that the two groups do not differ significantly in the proportion of correct answers made to questions concerning their reading habits. In general the percentage of failing students reporting use of psychologically correct reading habits does not differ significantly from the percentage so reporting from among successful students.

Habits of Concentration

Eight items of the questionnaire were concerned with different aspects of the student's attempt to concentrate on the material of his courses. In seven of the eight items the percentage giving the correct answer is larger for the failing groups, but in only one of these items is the difference between the two groups statistically significant. The mean difference in percentage for the eight items considered together is only 6, which is not statistically reliable. The findings indicate that correct habits of concentration are reported by failing and successful students in about the same proportions.

Attempts of Failing and Successful Students to Evaluate Critically and Analyze the Material of Their Courses

The percentage in each group giving the correct responses to the eleven items dealing with critical evaluation of course material is given in the table. In six of the items the larger percentage was found for the successful students, indicating a tendency for them to use methods of critical thinking more frequently than the students who are failing in their college studies. However, as was found for other aspects of study methods, the differences tended to be small. Three of the eleven differences are statistically significant; two of these favoring the failing students, the other favoring the successful students. If all the items are considered together and the mean percentages computed, the difference in the proportions reporting psychologically correct answers is only .2 per cent more for the failing than for the successful students. The facts support the conclusion that the two groups are not significantly different in terms of the proportion reporting correct answers to the items on critical evaluation.

Habits of Note Taking

Seven items covered different phases of note taking in regard to both textbook and lecture material. The table gives the results on this phase of study. For the most part the differences were small and statistically insignificant. In one question, the proportion of failing students giving the correct answer was significantly greater than the proportion of successful students. Results from the other items support the contention that there is no difference in the two groups in respect to the proportion of students reporting the use of correct habits of note taking.

Methods Used in Preparing for Examinations

Results from the seven items dealing with preparation for examinations are given in the table. Five of the differences favor the failing student. This might be expected from the fact that repeated failures would tend to stimulate the failing student to find specific ways of improving his performance on examinations. However, of the four statistically reliable differences, two favor the successful students. Apparently there is no wide discrepancy in the proportion of the two groups reporting the use of psychologically correct methods in preparing for examinations.

Methods of Reviewing

On five of the ten items dealing with methods of reviewing lecture and text materials the larger proportions of correct answers were given by the failing students and on the remaining five items by the successful students. Neither group is shown to be using the correct methods significantly more frequently than the other in nine of the ten items. Despite the fact that most of the standard crit-

ical ratios are larger than 1.5, the results do not warrant concluding that either group is superior to the other in methods of review.

Promptness in Preparing Assignments

Six items were concerned with the question of whether the two groups differed in the punctuality with which they completed the assignments of their courses. A larger percentage of successful students tend to fall behind in their assignments, and a larger percentage of the failing students prepare assignments ahead of schedule, but, in general, neither group shows marked superiority in the distribution of their study time.

General Summary of the Findings on the Seven Phases of Study Methods

The object of the present study was to determine if the difficulties encountered by failing students could be attributed to their utilization of unsound study habits. From the foregoing comparisons it is safe to conclude that, in general, the failing student reports using correct study habits as frequently as does the successful student. In fact, there is slight evidence that more failing students than successful students have developed sound study habits. No doubt many failing students could improve their academic achievement by improving their methods of study, but from the results of the present study it cannot be concluded that the difficulties encountered by failing students are in any large measure attributable to their utilization of unsound methods of study.

The Absolute Percentage of Failing Students Using Correct Study Habits

In the foregoing discussion the statement is made that probably many failing students could improve their scholarship by improving their study habits. This is based on the fact that the percentages reporting the use of correct study methods tended to fall between 25% and 75%, and that sometimes from 25% to 50% reported using methods which can be considered psychologically unsound. These facts can be shown to occur for items on which there is little disagreement as to the psychologically correct answer. For instance, in item 20 only 43% of failing students had developed the practice of glancing through a chapter, or of examining the main points in a chapter, before beginning a careful reading of it. There is little question but that either of these procedures of getting an overview of the whole chapter before reading it, is more sound psychologically than to start at the beginning of the chapter and carefully read it to the end. In item 16, in which the student is asked if he takes issue with the points of view presented in the lectures, only 34% of the failing students answered affirmatively. Again, in item 14, 34% of the failing students reported that they usually made their first review of their lecture notes just a day or so before an examination, rather than a short time after each lecture or during weekly reviews. In item 47, 37% of the failing students reported that they do not ask themselves questions about the material of an assignment while they are reading it.

The answers to the items just presented, together with similar answers to many other items, indicate that many of the students failing in their courses are not practicing sound study habits. No doubt improvement in their methods of study would be reflected in better grades. However, the same indictment can be made against the study habits of successful students, for large percentages of these students also are utilizing faulty study methods.

Limitations of the Results From the Questionnaire Method

In interpreting the foregoing results the limitations of the method should be kept in mind. There was no way of determining the accuracy of the individual answers, and therefore the reports may not describe the actual practices and habits of the students. It can be contended that the failing students would tend to falsify their reports to a greater degree than the successful students because of the pressure put upon them by their probationary status. The truth or falsity of this contention can never be established, and therefore it must be assumed that the successful and failing students were equally truthful in their answers to the items of the questionnaire.

The questionnaire method used here showed the number of failing and successful students reporting a given study habit. In most instances it did not obtain an estimate of the frequency with which the habit was used by the students, nor did it in any way get an estimate of the efficiency of the habit when it was used. Such limitations of the method

should be kept in mind when the results are being evaluated.

Concluding Statements

1. In general, the percentages of failing students reporting the use of correct study habits were as large as the percentages found among successful students.

2. In none of the larger phases of study methods investigated was the percentage of failing students using correct habits significantly below the percentage of successful students.

3. Large percentages of both failing and successful students fail to make use of many efficient study techniques.

Differences in Percentages of Failing and Successful Students Giving the Psychologically Correct Answers to Questions on Seven Phases of Study Habits

Topic	Number of Items	Average Percentage of Correct Answers for All Items		Distribution of Critical Ratios for Single Items According to Group Favored		
		Failing	Successful	Size of C. R.'s	Failing	Successful
Reading	20	61	58	.00–1.50	5	7
				1.51–1.99	2	2
				above 2.00	3	1
Concentration	8	40	34	0.00–1.50	4	1
				1.51–1.99	2	0
				above 2.00	1	0
Critical Evaluation	11	67	66	.00–1.50	2	3
				1.51–1.99	1	2
				above 2.00	2	1
Note Taking	7	54	50	.00–1.50	4	2
				1.51–1.99	0	0
				above 2.00	1	0
Preparation for Examinations	7	55	50	.00–1.50	0	0
				1.51–1.99	3	0
				above 2.00	2	2
Reviewing	10	51	50	.00–1.50	0	2
				1.51–199	4	8
				above 2.00	1	0
Promptness in Preparation	6	38	33	.00–1.50	1	1
				1.51–1.99	1	0
				above 2.00	2	1

THE NATURE OF SOME OF THE DIFFICULTIES OF STUDENTS FAILING THE FIRST TWO YEARS OF COLLEGE

C. W. Brown and P. V. Lofgren
University of California

Introduction. Numerous investigations have shown that ability to succeed in college is conditioned by many factors, not the least important of which are high school preparation, reading ability, and scholastic aptitude. Most of these studies have used representative groups of freshmen or sophomore college students. The present study attempts to isolate some of the characteristics of the relationships existing between the three variables mentioned and college success in a group heavily weighted with students who were failing to achieve an acceptable level of scholarship in their college work.

Selection of the Subjects. Failing students were selected from among those who were receiving special attention from the dean of their college because of their low scholarship. All of the subjects used completed their first four semesters of work at the University.[1] To these were added students who were comparable to the failing ones in terms of the number of semesters of work completed and the number of hours of course work essayed.

Measurement of the Variables. From the registrar's office the students' high school and college records were obtained and evaluated relative to the quality of the work performed. Letter grades were translated into numerical scores by assigning the following numbers to each grade: A:3, B:2, C:1, D:0, and E or F:—1.[2] Average grade point ratings were determined for each student for his four years of high school and his two years of college work. In addition, the high school achievement was evaluated separately for the following fields of subject matter: English and language, mathematics and natural science, social science, and biological science.

In studying the reading capacity of failing students it was desirable not only to learn their average reading ability, but to learn if any reading deficiencies existed for particular types of subject matter. None of the available reading tests for college students gave separate scores for the particular subject matters which were to be investigated, so it was necessary to construct new tests. The reading selections chosen for the different subject matters were: (1) for English and language, a selection on the exposition of the drama; (2) for mathematics and natural science, a selection on mechanical force; (3) for social science, a selection on Egyptian economic history; and (4) for biological science, a selection on respiration in lower organisms. These selections were taken from standard college texts which were in use or had been used recently in elementary courses in the respective fields of subject matter. The texts of the tests varied from 3 to 4 mimeographed pages. At the end of the reading selection in each test there was a set of from 20 to 24 multiple choice questions dealing with the material of the selection. The score for each subject matter test was the number of questions answered correctly. The score for the total reading test was the sum of the scores from the individual tests.

The time allowed for each subtest was 12 minutes. A preliminary buffer test was used to acquaint the student with the method to be followed in the remaining four tests. The student was required to read over a selection once before he began answering the questions; thereafter he could refer to the text again if he so desired. All of the details concerning the length of the selections, number of questions, time of testing, scoring, etc., had been worked out and standardized previous to the present study.

Scholastic aptitude was measured by means of the Ohio State University Psychological Examination, Form 19. It consisted of three subtests, an analogies test, an opposites test and a reading comprehension test. The subjects were given all the time they desired to complete the test, the average time required being about two and one-half hours.

Reliability of the Measures. The reliability of the grades of the individuals participating

[1] University of California, Berkeley.

[2] In evaluating the high school achievement, D, E, and F were given the value of zero.

in the present study was computed from the following formula:[1]

$$R = 1 - \frac{\sigma_{\Delta}^2}{\sigma_{\bar{x}_1}^2}$$

in which σ_{Δ}^2 is the mean, for N individuals, of $\sigma_{\Delta_1}^2$. The term $\sigma_{\Delta_1}^2$ is the sampling variance of the mean of the errors of measurement for any given individual for the full number of semesters he has been in school. The term $\sigma_{\bar{x}_1}^2$ represents the variance of the mean semester averages of N individuals. The reliability of the college grades was .88, and the reliability of the high school grades .75. These coefficients indicate reliabilities sufficiently high for the purposes of the present study.

The reliabilities of the separate reading tests were computed by the odd-even technique, using the Spearman–Brown prophesy formula. They varied from .79 to .90. The reliability of the total test was .92. The reliabilities of the subtests of the Ohio State University Psychological Examination, as reported by the author of the test are between .80 and .84. The reliability of the total test is .92.

Division of Subjects into Four Scholarship Groups. In order to study the three factors of high school achievement, reading ability, and scholastic aptitude, at different levels of college scholarship, the subjects were divided into four groups on the basis of their scholarship for the four semesters of college work. The grade point averages for the four groups, with the symbol assigned each group, follow: Group F1 consisted of the lowest failures, having a grade point average of .44 and meriting dismissal from the University; Group F2 included students who did not merit dismissal but who were deficient in scholarship with a grade point average of .85; Group C1 was an average group with a scholarship of 1.18 grade points; and Group C2 was a superior group with an average grade point standing of 1.98. The range of achievement represented in all four groups is from an average of —.30 grade points to 2.93 grade points.

In the comparisons to follow, the performance of Group C1, which is the C-average group, is used as a base, and the achievement of the other three groups is expressed as a deviation from the mean of the C1 group in terms of the sigma units of the C1 distribution.

Comparison of the Four Groups in Terms of the Average College Scholarship. When the college scholarship of Groups F1, F2, and C2 is expressed as relative achievement in terms of the performance of Group C1, we find the following results: the mean of Group F1 is 5.69 sigma units below the mean of Group C1, the mean of Group F2 is 2.54 sigma units below this mean, and the mean of Group C2 is 6.15 sigma units above the C1 mean. It is apparent that very large and significant differences in college scholarship exist between the mean achievements of adjacent groups.

Comparison of the Four Scholarship Groups in Terms of the Average Performance in the Three Experimental Variables. Fig. 1 represents graphically the relative achievement of the different groups in the three variables, using the performance of the C1 group as the point of reference. Each of the failing groups is seen to have a performance much below that of the C1 group, but the amount of the deficiency is somewhat different for the three variables. In the C2 group the greatest superiority is shown in high school achievement, and the least superiority in reading ability and scholastic aptitude.

Comparison of the Performance of the Four Groups in the Different Types of Subject Matter in High School. Fig. 2 presents a comparison of the relative performance in the four kinds of subject matter in high school of the four scholarship groups. The range of variation in relative performance in the different subjects is considerably greater in the F1 group than in either the F2 or C2 groups. The C2 group does not show uniform superiority to Group C1 in the four subject matters. The two subject matters best differentiating the failing groups from the C1 group are English and language and social science. The subjects of greatest differentiating capacity between Groups C1 and C2 are found in the mathematics and natural science category.

Comparison of the Performance of the Four Scholarship Groups in the Reading Tests Utilizing Different Types of Subject Matter. Fig. 3 gives a graphical representation of the relative performance of the four groups in the four subject matter reading tests. The range of variation is greatest in the F1 group and least in the C2 group. No one of the other

[1] From "Validation Against a Fallible Criterion" by E. F. Cureton. *Journal of Experimental Education*, I (1933), pp. 238–263.

groups agrees with the C1 group in their relative achievement in the different tests.

The natural science test highly differentiates the C2 group from the C1 group, but between the F2 and C1 groups, and between the F1 and F2 groups, the natural science test has the least differentiating capacity. The poorest performance of the C1 group in both high school achievement and the reading tests occurred in the field of mathematics and natural science. This somewhat general tendency toward weakness in this field would increase the difference between Groups C1 and C2 and decrease the differences between C1 and the failing groups, thus contributing to the conditions just noted. Another possible reason may be suggested. In the natural science reading test there is present less of the speed factor and more of the power factor than in any of the other time-limited tests. This condition might operate in this test to give the failing student a higher performance relative to the C1 group.

Comparison of the Performance of the Four Groups in the Subtests of Scholastic Aptitude. The relative performance of the four groups in the aptitude subtests is given in Fig. 4. Variation in relative performance is large in both failing groups. The two vocabulary tests, opposites and analogies, highly differentiate Groups F1 and F2. These subtests show greater differences between these two groups than any of the other individual measures in the three variables under study.

The Pattern of Achievement in the Four Types of Subject Matter in High School. In each of the variables under study the particular measurement used involved several factors, and it is important to note the variation in achievement among these factors in the four groups of subjects. The pattern of achievement in different types of subject matter in high school is given in Fig. 5. In order to facilitate comparison a line has been drawn connecting the averages of each group; no function is intended. It is seen that the relative achievement in the different subjects varies from one group to another. The pattern of Group C2 cannot be considered as typical, or to be expected from all levels of scholarship, because the higher averages are being depressed by the inclusion of large numbers of A grades, it being impossible for the individual to attain better than an A grade. No average in Group C1 is so affected; thus, the curve of this group probably represents the expected (normal) pattern of achievement.

The question arises as to whether the variations in the curve of Group C1 are due to chance, or reflect reliable differences in preparation in the different subjects. Standard critical ratios were computed for differences between adjacent points on each of the curves. They form the basis for the following statements. In Group C1 preparation in mathematics and natural sciences is significantly inferior to preparation in all other subjects; therefore, the initial rise in the curve of C1 reflects a real difference in quality of preparation. The change in trend between social science and biological science is not significant. The characteristics in the pattern of high school achievement in the C-average group may now be stated as follows: the preparation in mathematics and natural science is definitely inferior to the preparation in the other types of subject matter; the preparation in English and language, although superior to that in mathematics and natural science, is significantly inferior to the preparation in the two remaining subjects; and the difference in preparation in the social sciences and biological science subjects is not statistically significant.

The curve for Group F2 closely follows the form of the curve for C1. The first rise is statistically significant (C. R. 3.67). The difference between English and language and social science has a standard critical ratio of 1.87 which is strong evidence that a real difference in preparation in these two subjects might exist. The trend between social science and biological science varies slightly from that found in the curve for C1, but the change is not significant (C. R. .61). It would appear that in general, the pattern of preparation in high school subjects of Group F2 follows rather closely the pattern of Group C1.

In the curve of Group F1 the general trends again approach those of the C1 group. The first rise is statistically significant (C. R. 2.59). The second rise corresponds in general direction to the change in the other curves but is not significant (C. R. 1.27). The third portion of the curve, the rise between social science and biological science, although not significant (C. R. 1.49), is more pronounced than found for Group F2. With certain minor

variations the pattern of achievement in Group F1 corresponds to that of Group C1.

The present data are not sufficient for statistically determining the degree to which the variation in achievement in the different high school subjects conditions achievement in college; but regardless of what role the pattern of achievement plays in college study, it may be safely concluded that this role would be about the same for both failing and successful students. The principal differences between the several curves in Fig. 5 is not a difference in trends but a difference in levels of accomplishment. The failing groups differ from the successful groups in terms of lower levels of achievement in all fields of subject matter.

The Pattern of the Reading Performance of the Four Groups. Each of the reading tests was made long enough so that only a few perfect scores were attained. Also, sufficient time was allowed that no individual failed to reach and work on the questions. Every individual, then, performed within the range of the test, and the relative patterns of performance were not affected by factors similar to those operating in the evaluation of the high school preparation. Inasmuch as the raw scores on the various reading tests are not directly comparable, the average performance of each group was expressed as a deviation from the mean performance of Group C1, using the sigma of C1 as the unit of measurement. Fig. 6 presents graphically the pattern of achievement of the four groups. The pattern of C1 appears as a straight line at the value 0.00. Any divergence from a straight line in the curves of Group F1, F2, or C2, can be considered an indication of a variation in pattern from that of Group C1. The lines connecting the scores of a given group are drawn to facilitate comparison, no function being represented.

It is apparent that the pattern of performance of the F1 group does not follow closely the pattern of Group C1. The highest relative performance made by Group F1 was in the natural science test. Using the score in this test as a reference point, equality in relative achievement in the different reading tests can be represented by a straight line drawn parallel to the curve of C1. Any divergence from this line will indicate a deficiency in the relative reading achievement. It is obvious that Group F1 is particularly deficient in the reading of materials in English and language, and in social science. The tests in these fields were the two longest; they utilized vocabularies and dealt with concepts that were much simpler than was true for either the natural science or the biological science tests. From these facts the conclusion readily follows that the individuals of the F1 group are extremely slow readers. The F2 group shows a very marked deficiency in reading social science material. No doubt this latter deficiency is primarily due to the low rates of reading of these individuals.

The Pattern of Achievement in the Scholastic Aptitude Tests. Since the raw scores were not directly comparable, in order to plot the achievement in the different tests on the same coordinates, the performances of each group were expressed as deviations from the means of the C1 group. Fig. 7 presents the relative achievement of the four groups on the three parts of the aptitude test. As Group F1 made their highest relative performance on the opposites test, the scores on this test were selected as a point of reference. The pattern to be expected from successful college students can be considered to be the pattern of Group C1. It will be noted that the pattern of C2 diverges very little from this pattern, but that there are differences between the patterns of the failing groups and the C-average group.

From the data at hand it is not possible to determine if these variations in the pattern of scholastic aptitudes, as measured by the Ohio State University Psychological Test, contribute to differentiating failing and successful students in college. However, it is very probable that such differences of pattern as have been noted are too small to be of importance in conditioning college behavior. The general low level of performance in all of the aptitudes, rather than a difference in the pattern of achievement among the several aptitudes, seems to be the factor contributing to college failure.

Rate of Increase in Performance With Increase in College Scholarship. Some notion of the variations in the rate of increase in the performance in the three variables with increasing college scholarship can be gained from Fig. 1. In all three of the variables the most rapid change takes place between the levels of the F2 and C1 groups. This finding points to the existence of conditioning factors which vary in importance at different levels of college scholarship. It appears that the

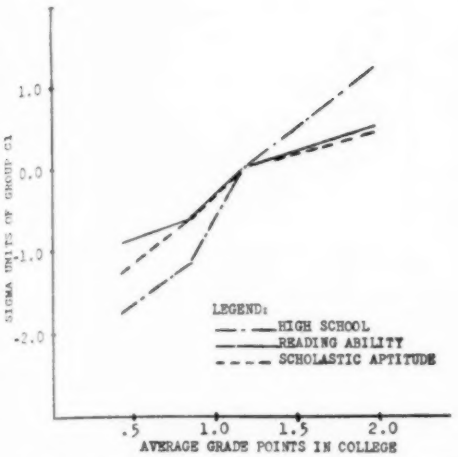


FIG. 1. COMPARISON OF THE FOUR SCHOLARSHIP GROUPS IN THE THREE VARIABLES USING MEAN OF GROUP C1 AS POINT OF REFERENCE.

power factors, functioning in high school achievement, still play an important role at the higher levels of college scholarship, as the slope of the curve for this variable still has a high value between C1 and C2.

Variations in the Rate of Increase in the Different Part Measures of the Three Variables With Increase in the Level of College Achievement. A re-examination of Fig. 2 will reveal an important difference in the rate of increase in the high school scholarship at different levels of college achievement. For all types of subject matter the curves change much more rapidly from the grade point level .85 to the level 1.18, than at any other levels. This change in rate is another indication of the depression of the average scores of Groups F1 and F2 in each of the four subject matters.

The true nature of the rate of change between Groups C1 and C2 for biological science, social science, and English and language cannot be determined because of the large number of perfect grades made in these subjects, especially by Group C2. This factor is not found in mathematics and natural science. The rate of change in achievement in this subject matter probably closely approximates the rate of change to be expected in the higher levels of college achievement.

The more rapid change in performance between F2 and C1 may be noted for each of the four time-limited reading tests by re-examining Fig. 3. This graph furnishes further proof that the more rapid change between F2 and C1, noted in preceding analyses, cannot be attributed to some subject matter factor. Whatever the cause of this difference, it seems to be operating in all of the types of subject matter analyzed in the present study.

From an examination of Fig. 4 it will be apparent that the change in scholastic aptitude is not constant with equal units of increase in college scholarship, and that the rate of increase varies with the different tests. Although the difference in college scholarship between Groups C1 and C2 is greater than between any other two adjacent groups, the increase in the associated scholastic aptitude is less than that found between other groups. A large number of perfect scores in Group

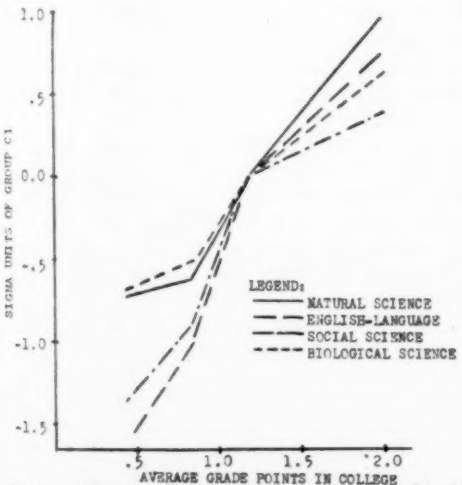


FIG. 2. COMPARISON OF THE FOUR SCHOLARSHIP GROUPS IN FOUR TYPES OF HIGH SCHOOL SUBJECT MATTER.

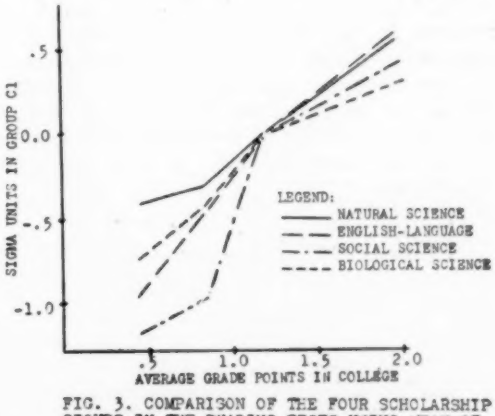


FIG. 3. COMPARISON OF THE FOUR SCHOLARSHIP GROUPS IN THE READING TESTS USING MEAN OF GROUP C1 AS POINT OF REFERENCE.

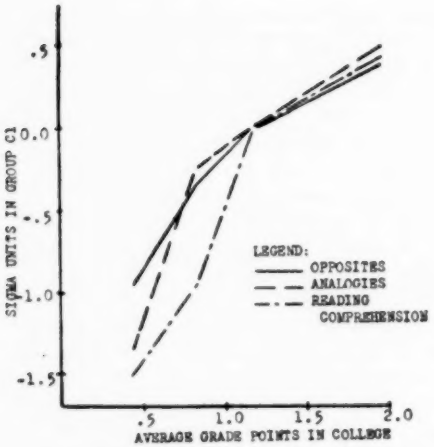


FIG. 4. COMPARISON OF THE FOUR SCHOLARSHIP GROUPS IN THE SUBTESTS OF SCHOLASTIC APTITUDE USING MEAN OF GROUP C1 AS POINT OF REFERENCE.

C_2 would account for this reduction of the slope, but since no large number of perfect scores on the tests was attained by any group, this explanation is not acceptable. A more plausible explanation is that the students of the C_2 group have approached closer to the maximal level of scholastic ability for their age. Increments of increase in test performance will get smaller the closer an individual approaches his maximal level. The approach to the limits of the student's scholastic ability, as measured by these tests, not the shortness of the tests, probably accounts for the reduction in the slope of the curve at the upper levels of achievement.

From the foregoing findings it is obvious that equal increments of increase in college scholarship are not associated with equal increments in performance on most of the part measures of the three variables being studied. Increase in college scholarship from a failing level to an average level is associated with the greatest increase in performance in most of the measures utilized in the present investigation.

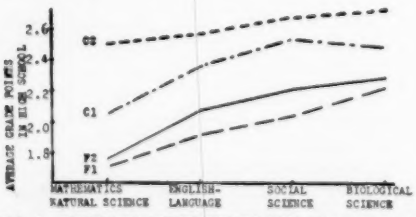


FIG. 5. PATTERN OF ACHIEVEMENT IN HIGH SCHOOL SUBJECT MATTERS

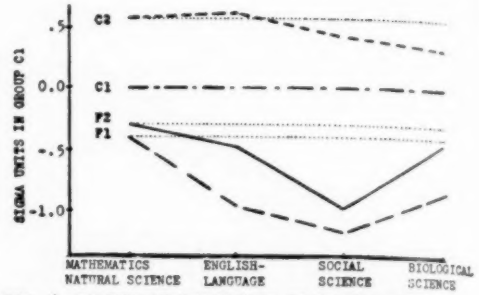


FIG. 6. PATTERN OF ACHIEVEMENT IN READING TESTS

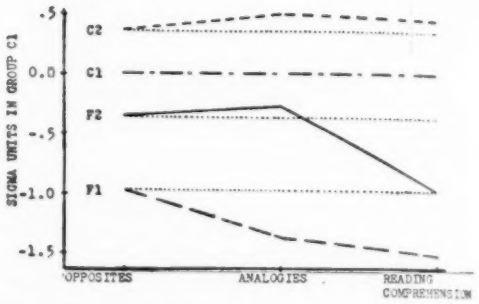


FIG. 7. PATTERN OF ACHIEVEMENT IN SCHOLASTIC APTITUDES

Concluding Statements. From the preceding analysis of the difficulties of students failing the first two years of college, the following conclusions may be drawn:

1. In the three variables of high school achievement, reading ability, and scholastic aptitude, the most inferior performance of failing students is in the variable of high school achievement.

2. Failing students achieve their highest relative performance in the variable of reading ability.

3. When high school achievement in the four fields of subject matter is examined, the failing students are found to give their lowest relative performance in English and language, and social science; and their highest relative performance in mathematics and natural science, and biological science.

4. In the reading tests utilizing different types of subject matter the failing groups achieved their lowest relative performance in reading English and language material, and their highest relative performance in reading material in natural science.

5. In the scholastic aptitude tests the lowest relative performance of the two fail-

ing groups was made in reading comprehension.

6. The range of variation in performance in the different part measures of each of the three principal variables tended to be greater in the failing groups than in either of the successful groups.

7. The pattern of achievement of the failing groups in the different high school subject matters did not vary greatly from the pattern of the successful groups.

8. In the reading tests and scholastic aptitude tests there was some evidence that the pattern of achievement of the failing groups differed from the pattern of the successful groups.

9. The most outstanding difference in the performance of the failing and successful groups was not a difference in pattern of achievement, but rather a difference in level of achievement, the performance of the failing groups being depressed in all of the measures studied.

10. The rate of increase in performance associated with increase in college scholarship was found to be higher for all measures between the levels of .85 and 1.18 grade points than between any other levels investigated. This most rapid change falls between the level of scholastic failure and the level of scholastic success.

A STUDY OF THE CAVD INTELLIGENCE SCALE AND ITS RESULTS ON THE DOCTOR'S LEVEL AT COLORADO STATE COLLEGE OF EDUCATION[1]

LORAINE BRUCE
Pampa Senior High School
Pampa, Texas

1. PROBLEM

The object of this study is to determine the homogeneity of the I. E. R. Intelligence Scale CAVD, Levels M, N, O, P, and Q, and its validity for the purpose of predicting achievement at the doctor's level in the Colorado State College of Education and to make comparisons of the abilities of students on the master's and doctor's levels in order to make a more adequate selection of candidates for work on the doctor's level. The CAVD Scale is made up of four series of tasks: Completions, arithmetical problems, vocabulary, and directions. The upper five levels M, N, O, P, and Q are for use on the upper college and graduate levels.

2. SUBJECTS AND TESTS

The students matriculating on the doctor's level from June, 1932 to August, 1938, were given the I. E. R. Intelligence Scale CAVD, Levels M, N, O, P, and Q, Form 4. In order to complete matriculation in 1936 and 1937, students on the doctor's level were required to take the American Council on Education Cooperative English Test and a battery of twelve education tests. English Usage and Spelling were the parts of the English Test which were used. The battery of education tests consists of tests in the following subjects: Techniques of curriculum making, elementary education, history of education, mental tests, personnel and guidance, philosophy of education, psychology of learning, research, school administration, secondary education, educational tests, and statistical methods. The tests were made up by various faculty members of Colorado State College of Education and were revised by the Personnel Department of the college. There are two equivalent forms of each test. All the items of each test are multiple-choice of four choices. The number of items in the various tests range from forty for research to fifty-seven for personnel and guidance. Data on each test were compiled from the group of ninety-one graduate students who first took the tests in the summer of 1936.

[1] Based on Field Study No. 2, Colorado State College of Education, Greeley, Colorado, 1940. This study is a continuation of the study, "The Relationship between Scores on the CAVD Intelligence Scale and Success in Graduate Work at Colorado State College of Education," which was published in the March, 1940, issue of the *Journal of Experimental Education*.

3. RESULTS

Of the students on the doctor's level who took the CAVD Intelligence Scale, 60 were in the graduate school at least two quarters and took a minimum of eight quarter-hours each quarter. The number of cases throughout the study is relatively small but no more were available; more reliable results might have been obtained from a larger number.

The reliability of teachers' marks on the basis of the graduate work done by this group of 60 students in one quarter is .433 $\pm$ .105 (SE). This is a relatively low reliability. In order to compute the reliability of teachers' marks and other correlations in which teachers' marks are involved, the point-hour ratio was found. The point-hour ratio was computed by dividing the total number of points by the total number of hours. The points were found by multiplying the number of hours of each course by the point values of the letter grades. The point value of "A" is 5; of "B" is 4; of "C", 3; of "D", 2; and of "F", 1.

The product-moment coefficient of correlation between scaled scores on the CAVD Intelligence Scale and the point-hour ratio based on the two quarters of the marks from which the reliability was computed is .521 $\pm$ .094. The coefficient of determination is .27; this coefficient gives the percentage of variance of teachers' marks which is associated with the CAVD Scale; the percentage of variance which is independent of the

CAVD Scale is .73. A half correction for attenuation was made in which the unreliability of teachers' marks was corrected, but in which no correction was made for the unreliability of the CAVD Test scores. This computation yielded a correlation of .670 between the CAVD Test scores and the theoretical true marks, which, rather than .521, is the measure of the true validity of the test when compared with a true criterion. The difference of the means of the point-hour ratio of the groups making above the 75th percentile (N = 49) and below the 25th percentile (N = 15) on the CAVD is .704. Since the standard error of the difference is .138, there is virtual certainty that there is a significant difference.

In addition to a comparison of the results of scaled scores on the CAVD Intelligence Scale and teachers' marks, a study was made of the relationship of each division of the CAVD Scale to teachers' marks. The product-moment coefficients of correlation (N = 46) obtained between the total point-hour ratio of teachers' marks and raw scores on the various divisions of the CAVD Scale are: (a) Completions, .491 ± .112; (b) arithmetic, .136 ± .145; (c) vocabulary, .297 ± .134; and (d) directions, .540 ± .105. These correlations show substantial relationships between completions and directions and teachers' marks, slight relationship between vocabulary and teachers' marks, and only negligible relationship between arithmetic and teachers' marks.

In order to complete matriculation in 1936 and 1937, students on the doctor's level were required to take the American Council on Education Cooperative English Test. The parts given were English Usage and Spelling. The product-moment coefficient (N = 54) between scaled scores on the CAVD Scale and scaled scores on the English Usage Test is .487 ± .104; that between scaled scores on the CAVD and spelling, an r of .610 ± .086. These r's show substantial relationships.

In 1936 and 1937, students on the doctor's level took Form A of the battery of twelve education tests. A product-moment coefficient of correlation (N = 48) of .691 ± .076 was obtained between scaled scores on the CAVD Scale and the average of weighted scores on the education tests. This correlation indicates an improvement of 49 per cent over the assignment of scores by mere chance and that the forecasting efficiency is 28 per cent. When this correlation was compared with the one for the students on the master's level, it was found that 97 times in 100, the correlation for students on the doctor's level will be the higher.

A comparison of abilities of students on the master's and doctor's levels was made by computing the differences between the means and standard deviations and standard error of each difference for the following measures: (a) Scaled scores on the CAVD Scale, (b) first quarter point-hour ratio, (c) second quarter point-hour ratio, (d) point-hour ratio based on the first two quarters, (e) scaled scores on the English Usage Test, (f) scaled scores on the spelling test, and (g) average of the total weighted scores on the battery of education tests.

The mean of the scaled scores on the CAVD Scale made by students on the doctor's level is 7.338 ± .884 higher than the mean made by students on the master's level; the difference of the standard deviations is .604 ± .625, with the SD of the masters the higher. Since the significance ratio (D/σ_D) for the difference of the means is 8.30 there is positive assurance that, on the basis of the CAVD Scale, the doctors are always superior to the masters. When the means of the three groups of point-hour ratios were subtracted, the means obtained from ratios made by the doctors were the higher in each case. The difference between the means of the point-hour ratios for the first quarter is .133 ± .058. The difference between the means of the point-hour ratios for the second quarter is .265 ± .072. And the difference between the means of the point-hour ratios based on the first two quarters is .221 ± .061. The significance ratio of 2.30 indicates that, on the basis of teachers' marks for the first quarter, doctors' students are not always superior to the masters. The significance ratios of 3.68 and 3.63, for marks for the second quarter and for a composite ratio of the first two quarters, respectively, indicate virtual certainty that, on the basis of these marks, the doctors are always superior to the masters. None of the differences between the standard deviations of marks are significant.

On the English Usage Test, the means and standard deviations of the masters are the higher. The difference between the means is 1.272 ± 1.319; and the difference between

the standard deviations is 1.056 ± .934. Neither of these differences is significant. The ratio (D/σ_D) of the differences between the means and the standard deviations on the spelling test indicate even less assurance. When the means of the scores on the education tests were subtracted, that of the doctors was higher by 4.105 ± .755. The significance ratio, 5.44, of this difference indicates with virtual certainty that, on the basis of the education tests, the doctors are superior to the masters. The standard deviation of scores on the education tests is higher by .643 ± .534, which is not a significant difference.

In order to study the homogeneity of the upper five levels of the CAVD Scale, all possible intercorrelations of its four parts were computed for 93 students on the doctor's level. From the correlations, $r_{CA} = .378$, $r_{CA.V} = .331$, $r_{CA.D} = .251$, and $r_{CA.VD} = .242$, it is evident that the correlation between C and A is affected more appreciably by D than by V. The net correlation between C and V is affected very little when A is eliminated, but D is evidently a factor in the correlation. This is shown by the following r's: $r_{CV} = .534$, $r_{CV.A} = .507$, $r_{CV.D} = .414$, and $r_{CV.AD} = .409$. The correlation between C and D is the highest of all, and indicates a substantial relationship between completions and directions. From the correlations, $r_{CD} = .553$, $r_{CD.A} = .491$, $r_{CD.V} = .442$, and $r_{CD.AV} = .386$, evidently both A and V affect the correlation between C and D. The correlation of .193 indicates only a negligible relationship between arithmetic and vocabulary. The partials, $r_{AV.C} = -.011$, $r_{AV.D} = .075$, and $r_{AV.CD} = -.033$, denote the further reduction of the correlation between A and V when C, D, and both C and D are eliminated. The correlation of .327 between A and D denotes low relationship. From the partials, $r_{AD.C} = .153$, $r_{AD.V} = .278$, and $r_{AD.CV} = .156$, C is a factor on which both A and D are dependent. The $r_{VD} = .392$ is not appreciably lowered when A is partialled out ($r_{VD.A} = .355$) but the elimination of C, which reduces the r to .137, indicates that vocabulary, directions, and completions are interdependent variables.

4. Conclusions

The I. E. R. Intelligence Scale CAVD, Levels M, N, O, P, and Q is a discriminating instrument for determining the range of abilities of graduate students on the doctor's level.

The correlation between the CAVD Scale and teachers' marks is relatively low; however, this correlation is consistent with the low reliability of teachers' marks on the doctor's level. The CAVD Scale is sufficiently discriminating to make it virtually certain that those students whose scores are above the 75th percentile on the CAVD will make higher marks than those students whose scores on the CAVD lie below the 25th percentile.

Although they evidently measure different abilities, substantial relationships were found between the CAVD Scale and the American Council on Education Cooperative English Usage Test and between the CAVD Scale and the American Council on Education Cooperative Spelling Test. There is a marked relationship between the CAVD Scale and the battery of education tests.

On the basis of comparisons of scores on the CAVD Scale, teachers' marks for the second quarter of graduate work, teachers' marks for the first two quarters of graduate work, and scores on the education tests, students on the doctor's level are always superior to students on the master's level. The significance ratios of 8.30, 5.44, 3.68, and 3.63 for the differences between the means of the scores made by students on the doctor's and master's levels on the CAVD Scale, the battery of education tests, marks for the second quarter, and composite marks for the first and second quarters, indicate the varying degrees of assurance that the doctors are superior to the masters' students. On the basis of the results obtained from these measures, a scale of values might be set up to be used in selecting candidates for the doctor's degree.

From the results of the intercorrelations of the four divisions of the CAVD Scale, the language tests (C, V, and D) are measures of different traits from those measured by the arithmetic test. Such correlation as appears between the language tests and arithmetic must be due to their common dependence upon reading comprehension. Evidently, A is measuring something very different from C and V and largely different from D. Since the divisions of the CAVD Scale do not overlap very much, each part makes a unique contribution to the whole.

ANALYSIS OF THE EFFECT UPON THE RELIABILITY COEFFICIENT OF CHANGES IN VARIABLES INVOLVED IN THE ESTIMATION OF TEST RELIABILITY

TEOBALDO CASANOVA
University of Puerto Rico

I. THE VARIANCE OF THE HALVES IN THE SPLIT-HALF METHOD

For computational purposes the reliability of a test may be defined as the correlation between two equivalent forms. For this purpose two forms are said to be equivalent when their means and variances are equal—a situation that rarely, if ever, occurs in actual practice. As neither the differences nor the significance of the differences between the statistics of one form and those of the other form are usually reported along with the reliability of the test, we have been used to think about this reliability without much concern about the equivalence of the forms correlated. When two forms are available their correlation is obtained through the use of their actual statistics. Otherwise the reliability is estimated by the split-half method, which owes its popularity to the fact that only one form is necessary for the purpose.

In this last method the correlation between the two halves of the available form is sometimes computed from actually observed data, and from this the correlation of the available form with a hypothetical form is estimated through the Spearman–Brown formula. Here, even when the standard deviations of the two halves of the available form differ, the four standard deviations—two of the available half and two of the hypothetical half—are implicitly assumed to be equal when the Spearman–Brown formula is applied. This contradiction may be avoided through the use of one of the formulas given below. By another method in which the reliability is estimated in terms of the variances of the sums and differences of the available halves, the four standard deviations are also assumed to be equal. It is obvious, therefore, that the use of different methods in the presence of actual disparity in the variance of the halves leads to different results which make difficult the comparison and transmutation of reliabilities computed under varying circumstances.

Let the subscripts *1* and *2* designate the scores on the two available halves of the test, and the subscripts *I* and *II* the scores on the corresponding halves of the hypothetical form. It is necessary to assume that $\sigma_1 = \sigma_I$ and that $\sigma_2 = \sigma_{II}$. But suppose that $\sigma_1 \pm \sigma_2$ and that $\sigma_I \pm \sigma_{II}$. It is also necessary to assume that the four intercorrelations are equal. Let each one of these be known as r, and let $X_1 + X_2 = S_1$, and $X_I + X_{II} = S_I$. Then,

$$r_{(X_1 + X_2)(X_I + X_{II})} = \frac{r(\sigma_1^2 + 2\sigma_1\sigma_2 + \sigma_2^2)}{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2},$$

or,

$$r_{S_1S_I} = \frac{r(\sigma_1 + \sigma_2)^2}{\sigma_{S_1}^2}. \qquad (1)$$

In practice r is the correlation between the two available halves, or the reliability of one-half of the test, and $r_{S_1S_I}$ is the estimated correlation of the available form with the hypothetical form, or the reliability of the whole test.

If

$$\sigma_1 = \sigma_2,$$

$$r_{S_1S_I} = \frac{2r}{1 + r}, \qquad (2)$$

which is the Spearman–Brown formula as used in the split-half method. Now S is the score on the sum of the two halves, or the score on the whole test. Let D be the score on the difference between the two halves. Then $S = 1 + 2$ and $D = 1 - 2$, and

$$\left.\begin{aligned}\sigma_S^2 &= \sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2.\\ \sigma_D^2 &= \sigma_1^2 - 2r\sigma_1\sigma_2 + \sigma_2^2.\end{aligned}\right\} \qquad (3)$$

Subtracting,

$$r = \frac{\sigma_S^2 - \sigma_D^2}{4\sigma_1\sigma_2}. \qquad (4)$$

Substituting this value in (1),

$$r_{S_1S_I} = \frac{\sigma_S^2 - \sigma_D^2}{\sigma_S^2} \cdot \frac{(\sigma_1 + \sigma_2)^2}{4\sigma_1\sigma_2}.$$

Since

$$(\sigma_1 + \sigma_2)^2 = (\sigma_1 - \sigma_2)^2 + 4\sigma_1\sigma_2,$$

$$r_{S_1S_1} = \left[1 - \frac{\sigma_D^2}{\sigma_S^2}\right]\left[1 + \frac{(\sigma_1 - \sigma_2)^2}{4\sigma_1\sigma_2}\right]. \quad (5)$$

The form of this may be changed to

$$r_{S_1S_1} = \left[1 - \frac{\sigma_D^2}{\sigma_S^2}\right]\left[\frac{1}{1 - \frac{(\sigma_1 - \sigma_2)^2}{(\sigma_1 + \sigma_2)^2}}\right]. \quad (5a)$$

Now, substituting (4) into (2),

$$r_{S_1S_1} = \frac{2(\sigma_S^2 - \sigma_D^2)}{\sigma_S^2 - \sigma_D^2 + 4\sigma_1\sigma_2},$$

and as from (3),

$$\sigma_S^2 + \sigma_D^2 = 2(\sigma_1^2 + \sigma_2^2) = 2(\sigma_1 - \sigma_2)^2 + 4\sigma_1\sigma_2,$$

$$r_{S_1S_1} = \frac{\sigma_S^2 - \sigma_D^2}{\sigma_S^2 - (\sigma_1 - \sigma_2)^2} \quad (6)$$

$$= (\sigma_S^2 - \sigma_D^2) \cdot \frac{1}{\sigma_S^2 - (\sigma_1 - \sigma_2)^2},$$

or,

$$r_{S_1S_1} = \left[1 - \frac{\sigma_D^2}{\sigma_S^2}\right]\left[\frac{1}{1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_S^2}}\right]. \quad (6a)$$

Formula (1), although never used, yields the most nearly correct estimate of the reliability of the test because it involves the least number of assumptions. It is easily seen that unless $\sigma_1 = \sigma_2$, its value is smaller than that given by (2), if its form is changed to

$$r_{S_1S_1} = \frac{r(\sigma_1 + \sigma_2)^2}{(\sigma_1 - \sigma_2)^2 + 2\sigma_1\sigma_2(1 + r)}. \quad (1a)$$

A common procedure is to find the correlation between the halves using the observed variances and then apply the Spearman–Brown formula. The value thus obtained is represented by formulas (6) and (6a), as may be inferred from their derivation. As the four variances are assumed to be equal in the derivation of the Spearman–Brown formula, there is a contradiction which may be avoided by using formulas (5) and (5a) which only involve the necessary assumptions that $\sigma_1 = \sigma_I$ and that $\sigma_2 = \sigma_{II}$. (5) and (5a) are different forms of (1).

Since

$$(\sigma_1 + \sigma_2)^2 > \sigma_S^2$$

unless $r = 1.00$, the value given by (6a) is larger than that given by (5a) unless $\sigma_1 = \sigma_2$, when the two values are equal.

If $\sigma_1 = \sigma_2$, each one of these formulas becomes

$$r = 1 - \frac{\sigma_D^2}{\sigma_S^2}. \quad (7)$$

This formula, derived by John Flanagan[1] is now frequently used, especially when the tests are scored in the scoring machine, because the *S*'s and the *D*'s are read directly in the dial. Its derivation assumes that the four variances are equal and its use, in case that $\sigma_1 \pm \sigma_2$ yields a smaller reliability than that obtained from either (5) or (6). But it is not necessary to find σ_D if the variance of one-half of the test is available. By using (3), (7) may be turned into

$$r_{S_1S_1} = \frac{4\sigma_1^2 r}{\sigma_S^2}. \quad (8)$$

But this requires a knowledge of r. From (2),

$$\frac{1 + r}{2r} = \frac{1}{r_{S_1S_1}}.$$

Then,

$$\frac{1}{r} = \frac{2}{r_{S_1S_1}} - 1.$$

Hence,

$$r = \frac{r_{S_1S_1}}{2 - r_{S_1S_1}}. \quad (9)$$

Substituting this value in (8),

$$r_{S_1S_1} = 2\left[1 - \frac{2\sigma_1^2}{\sigma_S^2}\right]. \quad (10)$$

in which σ_1^2 is the variance of one-half of the test, and σ_S^2 is the variance of the whole test. The formula is quite simple and the necessary data are easily available. The scores on one-half of the test are also read directly in the test scoring machine. It assumes that the four variances are equal, and its value is smaller than that given by either (5) or (6).

[1] Reported in Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educ. Rev.*, I (1939), p. 101.

II. The Correction for Guessing

1. *The Reliability of the Rights and the Reliability of the Wrongs.* The well known formula for correcting the scores in a multiple choice test for guessing is,

$$X = R - \frac{W}{n-1}, \tag{11}$$

in which X is the score corrected for guessing, R is the number of right answers, W is the number of wrong answers, and n is the number of choices. Dichotomous scoring is assumed throughout. Summing and dividing by N, the number of cases,

$$M_X = M_R - \frac{M_W}{n-1}, \tag{12}$$

where the M's are averages. If $N = \infty$, (12) becomes,

$$X_\infty = R_\infty - \frac{W_\infty}{(n-1)} \tag{13}$$

Here X_∞ is the true score after correction for guessing, R_∞ is the true score on the rights, and W_∞ is the true score on the wrongs. Squaring, summing, and dividing by the infinite population in (13),

$$\sigma_X^2 r_{X_1X_2} = \sigma_R^2 r_{R_1R_2} - 2r_{RW}\,\sigma_R\sqrt{r_{R_1R_2}}\,\frac{\sigma_W\sqrt{r_{W_1W_2}}}{n-1} + \frac{\sigma_W^2 r_{W_1W_2}}{(n-1)^2}. \tag{14}$$

Terms of the type $\sigma_X^2 r_{X_1X_2}$ represent the variance of true scores, and terms of the type $r_{X_1X_2}$ are reliabilities. Squaring, summing, and dividing by the population in (11),

$$\sigma_X^2 = \sigma_R^2 - 2r_{RW}\,\sigma_R\frac{\sigma_W}{n-1} + \frac{\sigma_W^2}{(n-1)^2}, \tag{15}$$

whence,

$$r_{X_1X_2} = \frac{\sigma_R^2 r_{R_1R_2} - 2r_{RW}\,\sigma_R\sqrt{r_{R_1R_2}}\,\frac{\sigma_W\sqrt{r_{W_1W_2}}}{n-1} + \frac{\sigma_W^2 r_{W_1W_2}}{(n-1)^2}}{\sigma_R^2 - 2r_{RW}\,\sigma_R\,\frac{\sigma_W}{n-1} + \frac{\sigma_W^2}{(n-1)^2}}. \tag{16}$$

This formula will not be generally useful in estimating the reliability of scores corrected for guessing because r_{RW} and $r_{W_1W_2}$ are not usually available, but it serves to show the effect of changes in the several variables upon the coefficient of reliability of scores corrected for guessing, $r_{X_1X_2}$.

Let

$$\frac{r_{W_1W_2}}{r_{R_1R_2}} = a^2.$$

Dividing both sides of (16) by $r_{R_1R_2}$,

$$\frac{r_{X_1X_2}}{r_{R_1R_2}} = \frac{\sigma_R^2 - 2r_{RW}\,\sigma_R\frac{\sigma_W}{n-1}a + \frac{\sigma_W^2}{(n-1)^2}a^2}{\sigma_R^2 - 2r_{RW}\,\sigma_R\frac{\sigma_W}{(n-1)} + \frac{\sigma_W^2}{(n-1)^2}}. \tag{17}$$

Now, r_{RW} is the correlation between the number of right answers and the number of wrong answers and it may be expected normally to be negative. As dichotomous scoring has been assumed throughout, if each one of the subjects taking the test attempts to answer all the items, $r_{RW} = -1.00$. Therefore, assuming that r_{RW} is negative, it is obvious that

$$r_{X_1X_2} > , = , \text{or} < r_{R_1R_2},$$

as

$$a > , = , \text{or} < 1.$$

In such a situation it may be concluded that if the reliability of the rights and that of the wrongs are equal, the correction for guessing is unnecessary. If the reliability of the rights is greater than that of the wrongs, the reliability of the corrected scores is smaller than that of the rights, and the contrary is true in the opposite case.

2. *The Variance of the Rights and the Variance of the Wrongs.* The reliability of the wrongs is not usually available, but the variance of the wrongs may be easily obtained. If the standard errors of measurement of the rights and of the wrongs are assumed to be equal, the reliability of the wrongs can be eliminated from equation (17). Doing so, and letting

$$\frac{\sigma_W}{\sigma_R} = \rho,$$

$$\rho^2 = \frac{1 - r_{R_1R_2}}{1 - r_{W_1W_2}}. \qquad (18)$$

Then,

$$1 - r_{W_1W_2} = \frac{1 - r_{R_1R_2}}{\rho^2},$$

or,

$$r_{W_1W_2} = 1 - \frac{1 - r_{R_1R_2}}{\rho^2},$$

and,

$$a^2 = \frac{r_{W_1W_2}}{r_{R_1R_2}} = \frac{1}{\rho^2} + \frac{1 - \frac{1}{\rho^2}}{r_{R_1R_2}}. \qquad (19)$$

Substituting this in (17) and dividing both members of the fraction on the right of (17) by σ_R^2,

$$\frac{r_{X_1X_2}}{r_{R_1R_2}} = \frac{1 - 2r_{RW}\frac{1}{n-1}\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right)^{1/2} + \frac{1}{(n-1)^2}\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right)}{1 - 2r_{RW}\frac{\rho}{n-1} + \frac{\rho^2}{(n-1)^2}}. \qquad (20)$$

Here it is evident that if $r_{R_1R_2} = 1$, $r_{X_1X_2} = r_{R_1R_2}$ whatever the value of the other variables; and from (17) it follows immediately that $r_{R_1R_2} = r_{W_1W_2} = r_{X_1X_2}$. Hence, if the assumptions made are true, tests whose scores uncorrected for guessing have proved to be highly reliable need not be corrected for guessing.

Equation (20) has been derived in order to study the dependence of the relative sizes of $r_{X_1X_2}$ and $r_{R_1R_2}$ on the value of ρ^2, which is the ratio of the variance of the wrongs to the variance of the rights. The last term in the numerator on the right of (20) can not be negative because it represents the ratio of two reliabilities. Subtracting from it the last term of the denominator on the same side,

$$\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right) - \rho^2 = \frac{\rho^2 - 1}{r_{R_1R_2}} - (\rho^2 - 1). \qquad (21)$$

It is easy to see here that if $r_{R_1R_2} < 1$, the term of the numerator $>$, $=$, or $<$ the term in the denominator as $\rho >$, $=$, or < 1. Continuing the assumption that r_{RW} is negative, if the other variables on the right of (20) remain constant, this leads directly to the conclusion that whether the size of $r_{X_1X_2}$ is greater, equal to or smaller than that of $r_{R_1R_2}$, depends respectively on whether the variance of the wrongs is greater, equal to or smaller than the variance of the rights. Thus the role of the variance of the rights is similar to that of the reliability of the rights, and the role of the variance of the wrongs is similar to that of the reliability of the wrongs, both of which would have been fair inferences to draw from the assumption expressed in equation (18). Since the presence of guessing in a test is indicated by the presence of the wrongs, at least to the extent to which the variance of the wrongs is a measure of the quantity and intensity (whether mild or wild) of guessing, the above conclusions are justified. One of these implies that if the variance of the wrongs is exaggerated in comparison with that of the rights, the reliability of the uncorrected scores (that of the rights) suffers a shrinkage in comparison with that of the corrected scores, thus requiring the application of the guessing correction due to excessive guessing.

The critical reader will not miss the implication that if we make a similar comparison of $r_{X_1X_2}$ with $r_{W_1W_2}$ by using the reciprocals of a and ρ in equations (17) and (20), the conclusions of this and the preceding subsection would still stand, but with $r_{W_1W_2}$ and $r_{R_1R_2}$ interchanged. This fact is readily disclosed by the symmetry of the right-hand side of (16) with respect to the two reliabilities and with respect to the two variances. However, the comparison of $r_{X_1X_2}$ and $r_{W_1W_2}$ is of no practical importance as the latter has never been used for any purpose.

Finally it is well to observe that if each one of the subjects attempts all the items of the tests, while dichotomous scoring is used, $r_{RW} = -1$, $\sigma_R = \sigma_W$, and therefore, $r_{X_1X_2} = r_{R_1R_2}$. This justifies the current practice of not correcting for guessing when all the subjects attempt all the items.

3. *The Correlation of the Rights With the Wrongs.* In order to study the effect of changes in r_{RW}, it is convenient to simplify the notation by putting A for the last term of the numerator of (20) and B for the last term of the denominator of the same equation. Thus (20) becomes,

$$\frac{r_{X_1X_2}}{r_{R_1R_2}} = \frac{1 - 2r_{RW}\sqrt{A} + A}{1 - 2r_{RW}\sqrt{B} + B}. \qquad (22)$$

Assuming that r_{RW} is negative, it is important to know now whether

$$\frac{\sqrt{A}}{1 + 2\sqrt{A} + A} > \text{or} < \frac{\sqrt{B}}{1 + 2\sqrt{B} + B},$$

where $A \pm B$ and $\rho \pm 1$; or whether

$$\frac{1}{\frac{1}{\sqrt{A}} + \sqrt{A} + 2} > \text{or} < \frac{1}{\frac{1}{\sqrt{B}} + \sqrt{B} + 2}.$$

But

$$\frac{1}{\sqrt{A}} + \sqrt{A}$$

has a minimum value of 2 when $A = 1$, and the same applies to B or to any other positive number. The value of this expression increases toward infinity both as A diminishes from 1 toward 0, and as A increases from 1 to ∞. It is most convenient to consider now only the case when $n = 2$. Then $B = \rho$.[2] It was proved through equation (21) that if $B > 1$, $A > B > 1$; and that if $B < 1$, $A < B < 1$. Therefore, unless $B = 1$, and in this case $r_{X_1X_2} = r_{R_1R_2}$, no matter what the value of r_{RW},

$$(A - 1)^2 > (B - 1)^2,$$

which means that A is more distant from 1 than B, whatever the signs of the differences. It is therefore proved that if $n = 2$ and $\rho \pm 1$,

$$\frac{2\sqrt{A}}{1 + 2\sqrt{A} + A} < \frac{2\sqrt{B}}{1 + 2\sqrt{B} + B}. \qquad (23)$$

It is deduced from this that if $r_{RW} = -1$, the proportion of the coefficient of r_{RW} in the numerator to the entire numerator of the right-hand side of (20) is smaller than the proportion of the coefficient of r_{RW} in the denominator to the entire denominator of the same equation. As r_{RW} rises from -1 to 0 in (20), if the other variables remain constant, the values of both members of the fraction decrease because r_{RW} is negative; but from inequality (23) it follows that the value of the numerator decreases less rapidly, which results in an increase in the ratio on the left of (20). In a similar manner it can be shown that as r_{RW} passes through 0 and continues its rise toward $+1$, the value of the ratio continues to increase until it reaches its maximum value when $r_{RW} = +1$.

The foregoing discussion has not included the point of whether $r_{X_1X_2} >$ or $< r_{R_1R_2}$. The ratio is determined by the values of a and ρ as indicated in the two preceding sections. Thus, in case that $r_{X_1X_2} > r_{R_1R_2}$, the increase of r_{RW} from -1 to 0 exaggerates the discrepancy; when $r_{X_1X_2} < r_{R_1R_2}$, the same increase in r_{RW} attenuates the discrepancy, and when $r_{X_1X_2} = r_{R_1R_2}$, changes in r_{RW} do not affect the ratio. Then if $a > 1$ and $\rho > 1$, so that $r_{X_1X_2} > r_{R_1R_2}$, a rise in r_{RW} from the value of -1 contributes to increase the difference and to make more necessary the use of the correction for guessing.

4. *The Variance of the Number of Items Attempted.* The correlation of the rights with the wrongs is not usually available, and for this reason equation (20) is mainly of theoretical importance. By using equations (3) the value of r_{RW} may be expressed in terms of the variances of sums and differences, which are more readily obtained, especially when the tests are scored in the test scoring machine. Changing 1 and 2 for W and R, and changing S for T we have from (4) that

$$r_{RW} = \frac{\sigma_T^2 - \sigma_D^2}{4\sigma_R\sigma_W}, \qquad (4a)$$

where $T = R + W$, and $D = R - W$; that is, T is the total number of items attempted by one subject and D is the difference for each subject between his score on the rights and his score on the wrongs. Substituting this in (20),

$$\frac{r_{X_1X_2}}{r_{R_1R_2}} = \frac{1 - \frac{\sigma_T^2 - \sigma_D^2}{2\sigma_R\sigma_W(n-1)}\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right) + \frac{1}{(n-1)^2}\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right)}{1 - \frac{\sigma_T^2 - \sigma_D^2}{2\sigma_R\sigma_W(n-1)}\rho + \frac{\rho^2}{(n-1)^2}}. \qquad (24)$$

Bearing in mind the changed notation, it is evident from (3) that as r_{RW} rises from -1 to $+1$ the value of σ_T increases while the value of σ_D decreases continuously for fixed values of σ_R and σ_W. From the conclusions and assumptions of the preceding section it is inferred that an increase in σ_T produces an increase in the ratio on the left of (24). This is quite in agreement with current notions, according to which it is the fact that all the subjects do not attempt all the items that requires the use of the guessing correction. Since dichotomous scoring is assumed, σ_T^2 is the variance of the number of items attempted as well as the variance of the number of items missed. But while it is true that an increase in the value of σ_T increases $r_{X_1X_2}$ in relation to $r_{R_1R_2}$ if the other variables remain constant, it is also true that this increase tends to make the two reliabilities equal if $\rho < 1$. If $\rho > 1$, an increase in σ_T further increases the value of $r_{X_1X_2}$ over that of $r_{R_1R_2}$, while if $\rho = 1$, changes in σ_T do not affect the equality of $r_{X_1X_2}$ and $r_{R_1R_2}$ if the assumptions made are correct. It is to be noted here again that if all the subjects attempt all the items, $r_{RW} = -1$, and from (3) it is evident that $\sigma_R = \sigma_W$. Then $\sigma_T = 0$, and as it is well known, the guessing correction is unnecessary.

5. *The Number of Choices.* If $\rho = 1$, $r_{X_1X_2} = r_{R_1R_2}$, whatever the value of n. If $\rho < 1$, then, returning to subsection 3, $A < B < 1$, and inequality (23) still holds with all its consequences. An increase in n would accelerate the rate at which a rising r_{RW} and a rising σ_T increase the value of $r_{X_1X_2}$ over that of $r_{R_1R_2}$. But since in these circumstances $r_{X_1X_2} < r_{R_1R_2}$, the effect of an increase in n is to diminish the difference between $r_{X_1X_2}$ and $r_{R_1R_2}$.

If $\rho > 1$, B, which is equal to $\dfrac{\rho^2}{(n-1)^2}$, may be $>$, $=$, or <1. If $B > 1$, $A > B > 1$, and inequality (23) holds again. Since $r_{X_1X_2} > r_{R_1R_2}$, an increase in n would tend to exaggerate the inequality. However, considering that n is stated in whole numbers, and that this section is dealing with values of $n =$ or > 3, the cases where $n = 2$ having already been treated in the preceding sections, it is doubtful whether n can be increased over the value of 3 to any considerable extent while B remains greater than 1, or that is, when $\rho > (n-1)$.

If $B < 1$ while $\rho > 1$, we know that

$$\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right) > \rho^2 > 1,$$

but A, which is equal to the quantity on the left divided by $(n-1)^2$, may be $>$, $=$, or < 1. If

$$\left(1 + \frac{\rho^2 - 1}{r_{R_1R_2}}\right) < (n-1)^2; \qquad (25)$$

$A < 1$ and $B < A < 1$. Here inequality (23) is reversed, and the effect of an increase in r_{RW} and σ_T is now to diminish the ratio of $r_{X_1X_2}$ to $r_{R_1R_2}$, while n remains large enough so that (25) holds true. As now $r_{X_1X_2} > r_{R_1R_2}$ because $\rho > 1$, it is seen that a further increase in n serves to exaggerate the discrepancy in A and B and thus to accelerate the effect of a rising r_{RW} and a rising σ_T in levelling down the difference between $r_{X_1X_2}$ and $r_{R_1R_2}$.

If $A =$ or > 1 while $B < 1$ and $\rho > 1$, inequality (25) becomes an equation or it is reversed. In case it is reversed, $B < 1 < A$, and it is necessary to know which one of the two variables is more distant from 1, in order to know whether inequality (23) or its reversed form hold. However, this case is of no great practical importance because, as we are dealing with values of $n =$ or > 3, we are very unlikely to find a situation where inequality (25) does not hold. In conclusion, it may be said that, with the exception of some of the rare cases in which ρ has an exceedingly large value, which may upset inequality (25), a rise in the value of n has a tendency to equalize $r_{X_1X_2}$ and $r_{R_1R_2}$, which actually happens when $n = \infty$.

6. *Determinations in Terms of the Halves.* The split-half method enables us to make formulas in terms of more easily available data for the estimation of the reliability of scores corrected for guessing, without going through the usual procedure. Returning to the notation of Section I, and preserving the meaning of R, W, and X in Section II, we have from formula (10),

$$r_{X_{s_1}X_{s_I}} = 2\left[1 - \frac{2\left(\sigma_{R_1}^2 - 2r_{R_1W_1}\sigma_{R_1}\dfrac{\sigma_{W_1}}{(n-1)} + \dfrac{\sigma_{W_1}^2}{(n-1)^2}\right)}{\sigma_{R_s}^2 - 2r_{R_sW_s}\sigma_{R_s}\dfrac{\sigma_{W_s}}{n-1} + \dfrac{\sigma_{W_s}^2}{(n-1)^2}}\right], \qquad (26)$$

where the values of σ_1 and σ_8 in (10) have been substituted from (15). If the two variances are equal in (3),

$$\sigma_{R_8}^2 = 2\sigma_{R_1}^2(1 + r_{R_1R_2}), \tag{27}$$

or, by this and (9),

$$\sigma_{R_1}^2 = \frac{\sigma_{R_8}^2(2 - r_{R_{S1}R_{SI}})}{4}. \tag{28}$$

In a similar manner,

$$\sigma_{W_1}^2 = \frac{\sigma_{W_8}^2(2 - r_{W_{S1}W_{SI}})}{4}. \tag{28a}$$

This equation is almost equivalent to (16), although the value of the reliability coefficient is slightly higher here if r_{RW} is negative. It could have been derived directly from (11) in terms of obtained scores, while (16) was derived from (12) in terms of true scores, on account of which the latter is more accurate and was preferred by the writer for the analysis made in the preceding section. From (3),

$$2r_{R_8W_8}\sigma_{R_8}\frac{\sigma_{W_8}}{(n-1)} = \frac{\sigma_{T_8}^2 - (\sigma_{R_8}^2 + \sigma_{W_8}^2)}{(n-1)}.$$

Substituting from here and from (10) into (30),

$$r_{X_{S1}X_{SI}} = \frac{2(\sigma_{R_8}^2 - 2\sigma_{R_1}^2) - \frac{\sigma_{T_8}^2 - (\sigma_{R_8}^2 + \sigma_{W_8}^2)}{(n-1)} + \frac{2(\sigma_{W_8}^2 - 2\sigma_{W_1}^2)}{(n-1)^2}}{\sigma_{R_8}^2 - \frac{\sigma_{T_8}^2 - (\sigma_{R_8}^2 + \sigma_{W_8}^2)}{(n-1)} + \frac{\sigma_{W_8}^2}{(n-1)^2}}$$

$$= \frac{\sigma_{R_8}^2\left(1 + \frac{1}{n-1}\right) + (\sigma_{R_8}^2 - 4\sigma_{R_1}^2) - \frac{\sigma_{T_8}^2}{(n-1)} + \sigma_{W_8}^2\left(\frac{1}{n-1} + \frac{1}{(n-1)^2}\right) + \frac{\sigma_{W_8}^2 - 4\sigma_{W_1}^2}{(n-1)^2}}{\sigma_{R_8}^2\left(1 + \frac{1}{(n-1)}\right) - \frac{\sigma_{T_8}^2}{n-1} + \sigma_{W_8}^2\left(\frac{1}{n-1} + \frac{1}{(n-1)^2}\right)},$$

or,

$$r_{X_{S1}X_{SI}} = 1 - \frac{(4\sigma_{R_1}^2 - \sigma_{R_8}^2) + \frac{1}{(n-1)^2}(4\sigma_{W_1}^2 - \sigma_{W_8}^2)}{\sigma_{R_8}^2\frac{n}{n-1} - \frac{\sigma_{T_8}^2}{n-1} + \sigma_{W_8}^2\left[\frac{n}{(n-1)^2}\right]}. \tag{31}$$

Continuing the Spearman–Brown assumptions of Section I,

$$r_{R_8W_8} = \frac{4r_{R_1W_1}\sigma_{R_1}\sigma_{W_1}}{\sigma_{R_8}\sigma_{W_8}},$$

and substituting the values of σ_{R_8} and σ_{W_8} as given in (27)

$$r_{R_1W_1} = \frac{1}{2}r_{R_8W_8}\sqrt{1 + r_{R_1R_2}}\sqrt{1 + r_{W_1W_2}}. \tag{29}$$

Substituting into (26) from (27), (28), (28a), and (29),

$$r_{X_{S1}X_{SI}} = 2 - \frac{4\left[\frac{\sigma_{R_8}^2(2 - r_{R_{S1}R_{SI}})}{4} - \frac{1}{2}r_{R_8W_8}\sigma_{R_8}\frac{\sigma_{W_8}}{(n-1)} + \frac{\sigma_{W_8}^2(2 - r_{W_{S1}W_{SI}})}{(n-1)^2}\right]}{\sigma_{R_8}^2 - 2r_{R_8W_8}\sigma_{R_8}\frac{\sigma_{W_8}}{(n-1)} + \frac{\sigma_{W_8}^2}{(n-1)^2}},$$

or,

$$r_{X_{S1}X_{SI}} = \frac{\sigma_{R_8}^2 r_{R_{S1}R_{SI}} - 2r_{R_8W_8}\sigma_{R_8}\frac{\sigma_{W_8}}{n-1} + \frac{\sigma_{W_8}^2}{(n-1)^2}r_{W_{S1}W_{SI}}}{\sigma_{R_8}^2 - 2r_{R_8W_8}\sigma_{R_8}\frac{\sigma_{W_8}}{(n-1)} + \frac{\sigma_{W_8}^2}{(n-1)^2}}. \tag{30}$$

This is a simple formula which does not contain correlations or reliabilities, but it only gives a rough estimate of $r_{X_{S1}X_{SI}}$ in case the assumptions made are not strictly fulfilled. For a closer estimate, (24) can be used.

III. The Effect of Calling Zero All Negatives Scores

Let X_a designate a score of zero or a score above zero, X_b a score below zero after cor-

rection for guessing, and X any score, whether zero, above zero, or below zero. Then

$$\Sigma X = \Sigma X_a + \Sigma X_b, \tag{32}$$

and

$$\sigma_X^2 = \frac{\Sigma X_a^2 + \Sigma X_b^2}{N} - \frac{(\Sigma X_a + \Sigma X_b)^2}{N^2}. \tag{33}$$

This is the complete variance of the scores on a test. Let σ_{Xe}^2 represent the variance obtained when all the X_b scores are called zero. Now all the b's vanish, and since N, the number of cases is not reduced by calling zero the negative scores,

$$\sigma_{Xe}^2 = \frac{\Sigma X_a^2}{N} - \frac{(\Sigma X_a)^2}{N^2},$$

and,

$$\sigma_{Xe}^2 = \sigma_X^2 - \left[\frac{\Sigma X_b^2}{N} - \frac{(\Sigma X_b)^2}{N^2} - \frac{2\Sigma X_a \Sigma X_b}{N^2}\right]. \tag{34}$$

If N_b is the number of b scores,

$$\sigma_b^2 = \frac{\Sigma X_b^2}{N_b} - \frac{(\Sigma X_b)^2}{N_b^2}.$$

Therefore,

$$\frac{\Sigma X_b^2}{N} = \frac{N_b}{N}\left[\sigma_b^2 + \frac{(\Sigma X_b)^2}{N_b^2}\right].$$

Substituting in (34), and letting N_a be the number of a scores,

$$\sigma_{Xe}^2 = \sigma_X^2 - \frac{N_b}{N}\left[\sigma_b^2 + \frac{(\Sigma X_b)^2}{N_b^2}\right] + \frac{(\Sigma X_b)^2}{N^2} + \frac{2N_a M_a N_b M_b}{N^2}$$

$$= \sigma_X^2 - \frac{N_b}{N}\left[\sigma_b^2 + \left(M_b^2 - \frac{N_b}{N} M_b^2\right) - \frac{2N_a}{N} M_a M_b\right].$$

Since $N = N_a + N_b$,

$$\sigma_{Xe}^2 = \sigma_X^2 - \frac{N_b}{N}\left[\sigma_b^2 + \frac{N_a}{N}(M_b^2 - 2M_a M_b)\right]. \tag{35}$$

If the measures are the sums of the scores on two equivalent halves of a test, all that is necessary to express the relationship is to change the notation. Thus,

$$\sigma_{Se}^2 = \sigma_S^2 - \frac{N_b}{N}\left[\sigma_{Sb}^2 + \frac{N_a}{N}(M_{Sb}^2 - 2M_{Sa}M_{Sb})\right]. \tag{35a}$$

However, if the measures used are the differences between the two halves, and if these are assumed to have equal means and variances, the means vanish, and

$$\sigma_{De}^2 = \sigma_D^2 - \frac{N_b}{N}\sigma_{Db}^2.$$

Substituting into (7) from (35a) and (36)

$$r_{1Ise} = \frac{\sigma_S^2 - \sigma_D^2 - \frac{N_b}{N}\left[(\sigma_{Sb}^2 - \sigma_{Db}^2) + \frac{N_a}{N}(M_{Sb}^2 - 2M_{Sa}M_{Sb})\right]}{\sigma_S^2 - \frac{N_b}{N}\left[\sigma_{Sb}^2 + \frac{N_a}{N}(M_{Sb}^2 - 2M_{Sa}M_{Sb})\right]},$$

in which r_{1Ise} is the reliability of the test after all negative scores have been called zero.

From (7),

$$\sigma_S^2 - \sigma_D^2 = r_{1Is}\,\sigma_S^2,$$

and,

$$\sigma_{Sb}^2 - \sigma_{Db}^2 = r_{1Ib}\,\sigma_{Sb}^2.$$

Hence,

$$r_{1Ise} = \frac{r_{1Is}\,\sigma_S^2 - \frac{N_b}{N}\left[r_{1Ib}\,\sigma_{Sb}^2 + \frac{N_a}{N}(M_{Sb}^2 - 2M_{Sa}M_{Sb})\right]}{\sigma_S^2 - \frac{N_b}{N}\left[\sigma_{Sb}^2 + \frac{N_a}{N}(M_{Sb}^2 - 2M_{Sa}M_{Sb})\right]}. \tag{37}$$

r_{1Is} is the reliability of the test computed from all the scores, and r_{1Ib} is the reliability of the scores below zero after correction for guessing. The fraction on the right is a proper fraction and the quantity in brackets is positive if r_{1Ib} is positive, because M_{Sb},

which is the mean of the negative scores for the whole test is negative, while M_{Sa}, the mean for the scores of zero or above is positive. Granting that r_{1Ib} is positive and large, it may be seen in (37) that $r_{1Ise} < r_{1Is}$, and that an increase in the quantity in brackets will increase the difference between r_{1Ise} and r_{1Is} by diminishing r_{1Ise}. Other changes that will diminish r_{1Ise} are increases in N_b the number of negative scores, σ_{Sb} and M_{Sb}. As M_{Sb} is negative, it means that the further below zero the location of M_{Sb}, the smaller r_{1Ise} in relation to r_{1Is}. In conclusion, it may be said that a considerable loss of reliability in the scores corrected for guessing may be expected after those scores that fall below zero have been eliminated, if the latter are many in number and spread far below zero.

IV. The Variance of the Items

Kuder and Richardson[1] have given the following formulas for estimating the reliability of a test in terms of the variance of the items:

$$r_{11} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\,\overline{pq}}{\sigma_t^2}, \qquad (38)$$

$$r_{11} = \frac{n}{n-1} \frac{\sigma_t^2 - n\,\bar{p}\,\bar{q}}{\sigma_t^2}, \qquad (39)$$

in which n is the number of items, σ_t^2 is the test variance, $\overline{pq}$ is the average item variance, $\bar{p}$ is the average p and $\bar{q}$ is the average q, the proportion of persons missing one item; and $p + q = 1$. (39) is obtained from (38) by assuming that all the item variances are equal. But this never happens in test construction so that (39) is subject to an error, which may be of considerable size if there are wide differences among the item variances. If the difference between the two values is D, subtracting the second from the first,

$$D = \frac{n^2}{(n-1)\,\sigma_t^2}(\bar{p}\,\bar{q} - \overline{pq}). \qquad (40)$$

But

$$\bar{p}\,\bar{q} - \overline{pq} = -r_{pq}\,\sigma_p\,\sigma_q.$$

As all the p's are not equal, $r_{pq} = -1$. Also, $\sigma_p = \sigma_q$, assuming dichotomous scoring. Then,[2]

$$\bar{p}\,\bar{q} - \overline{pq} = \sigma_p^2. \qquad (2)$$

[1] Kuder, G. F., and Richardson, M. W. "The theory of the estimation of test reliability." *Psychometrika*, III (1937), p. 158.

Since σ_p is positive it is evident that the value given by (38) is always larger than that given by (39), and

$$D = \frac{n^2\,\sigma_p^2}{(n-1)\,\sigma_t^2}.$$

By adding this to the right-hand side of (39), the value given by (38), which is the more nearly correct value, is obtained.

Doing so,

$$r_{11} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\,\bar{p}\,\bar{q}}{\sigma_t^2} + \frac{n^2\,\sigma_p^2}{(n-1)\,\sigma_t^2},$$

or,

$$r_{11} = \frac{n}{n-1}\left[1 + \frac{n\,(\sigma_p^2 - \bar{p}\,\bar{q})}{\sigma_t^2}\right]. \qquad (41)$$

But,

$$\sigma_p^2 = \frac{\Sigma\,p^2}{n} - \bar{p}^2,$$

and since $p + q = 1$,

$$\bar{p}\,\bar{q} = \bar{p}\,(1 - \bar{p}) = \bar{p} - \bar{p}^2.$$

Substituting these in (41),

$$r_{11} = \frac{n}{n-1}\left[1 + \frac{n\left(\frac{\Sigma\,p^2}{n} - \bar{p}\right)}{\sigma_t^2}\right],$$

or,

$$r_{11} = \frac{n}{n-1}\left[1 - \frac{\Sigma\,p - \Sigma\,p^2}{\sigma_t^2}\right]. \qquad (41a)$$

This simple formula, which can be obtained directly from (38), is more convenient than (38) for computing the value of r_{11} because it does not involve the products. It shows that $\bar{p}\,\bar{q} > \sigma_p^2$, as was to be expected in (41) from the fact that $r_{11} < 1$. The term $\bar{p}\,\bar{q}$ can be changed in (41) without changing σ_p. This may be accomplished by adding a constant to each p. If the distribution of p's is symmetrical with respect to $\bar{p}$, it is evident that as the distribution of p's is moved up and down the range, $\bar{p}\,\bar{q}$ varies, and reaches its maximum when $\bar{p} = .50$. If the entire range of p's is less than .50, it is possible to have the whole distribution of p's either above or below .50. In either case $\bar{p}\,\bar{q}$ would have a low value in comparison with the max-

[2] This σ_p should not be confused with the standard error of the proportion in one category of a Bernoulli distribution.

imum value of .25. This means that, according to formula (41), for a given σ_p, and if it is possible for the other variables to remain constant, the value given by (41) is a minimum when the mean of the distribution of p's, which has been assumed to be symmetrical, is at .50. That is, a distribution of p's with a mean of .50, which is the so-called median difficulty of an item, produces a less reliable test than a distribution of p's with an average difficulty near 1 or near 0. This may be interpreted as being in conflict with Symonds' well known findings[3] on the relation between item difficulty and reliability, but the point noted here is different; and if formula (41) is correct, the present conclusion is unavoidable. It seems hardly necessary to comment here that an increase in either σ_t or n will result in an increase of r_{11}.

[3] Symonds, P. M. "Choice of items for a test on the basis of difficulty." *Journal of Educational Psychology*, VII (1929), pp. 481–493.

ITEM VALIDITIES OF THE OTIS SELF-ADMINISTERING TESTS OF MENTAL ABILITY FOR A COLLEGE POPULATION

WILLIAM R. CROOKS
Suffolk University
and
LEONARD W. FERGUSON
University of Connecticut

This study was undertaken in order to provide data concerning the difficulty and validity of the items on the Otis test for a college population comparable to those which Chapanis[1] has furnished for adults. The population from which the results were secured was the student body at the University of Connecticut. All tests were administered as laboratory exercises to the students (mostly freshmen and sophomores) in elementary psychology. A description of the sample is provided in Table I. The data suggest that Form B is slightly more difficult from Form A. (The records for Form A were randomly selected from 1000 test records, but records for Form B represent all those available at the time of study.) It also appears that the test is either slightly more difficult for women or else that women are somewhat inferior to men in intelligence. Since so much evidence has been presented in refutation of the second alternative, we are left with only the first, if the differences are interpreted as significant.

TABLE I

DESCRIPTION OF THE SAMPLE

	N*	Mean	Range	S D	Sex Diff.
Form A					
Men	100	60.2	37–73	8.0	2.5 ± .75
Women	100	57.7	40–72	7.7	
Form B					
Men	137	55.0	35–69	7.7	
Women	100	52.8	22–70	10.1	2.2 ± .81

* There is no overlap among these groups; those students who took Form A did not take Form B.

In Table II are presented data showing the relative difficulties (per cent of correct responses) and relative validities (biserial correlations) of all items in Form A of the Otis test. Presented for comparison with our own figures are those previously reported by Chapanis. Examination of the data in this table reveals the following facts:

1. The median difference for men between the per cent of adults and the per cent of students who got each item correct is .08 (Range .00 to .25).

2. The corresponding median difference for women is .09 (Range .01 to .35).

3. The median sex difference for Chapanis' data in per cent of items correct is .04 (Range .00 to .23).

4. The corresponding median for Crooks and Ferguson's data is .03 (Range .00 to .29).

5. The median difficulty of the items on the basis of Chapanis' data is .86 for women and .79 for men.

6. The corresponding median on the basis of Crooks and Ferguson's data is .86 for women and .87 for men.

7. The median item validity on the basis of Chapanis' data is .61 for women and .62 for men.

8. The corresponding median on the basis of Crooks and Ferguson's data is .48 for women and .46 for men.

9. There are (according to the writers' data) 18 items for college men and 23 items for college women which have non-significant validities. Nine of these are common to both men and women.

[1] Chapanis, A. "A note on the validity and difficulty of items in Form A of the Otis Self-Administering Tests of Mental Ability," *Journal of Experimental Education*, V (1937), pp. 246–248.

TABLE II

VALIDITIES AND DIFFICULTIES OF THE ITEMS IN (HIGHER) FORM A OF THE OTIS SELF-ADMINISTERING TEST FOR A COLLEGE AND AN ADULT POPULATION[2]

Item No.	Crooks-Ferguson Men r	Crooks-Ferguson Men Per Cent Correct	Crooks-Ferguson Women r	Crooks-Ferguson Women Per Cent Correct	Chapanis Men r	Chapanis Men Per Cent Correct	Chapanis Women r	Chapanis Women Per Cent Correct
1	....	100	....	100	.93	96	.99	97
2	.27	96	.21	92	.77	89	.45	83
3	.26	90	.31	90	.25	86	.38	86
4	.25	98	.27	97	.76	93	.35	89
5	.47	93	.06	92	.23	85	.46	88
6	.37	87	.28	88	.39	86	.30	84
7	—.34	97	.11	95	.50	84	.47	84
8	.29	89	.21	92	.54	79	.42	86
9	—.13	99	....	100	.96	99	.62	99
10	....	100	.26	98	.74	97	.51	95
11	.34	99	.71	99	.31	92	.61	91
12	—.02	98	.26	98	.68	93	.88	93
13	.92	98	.94	97	.50	92	.74	91
14	.14	94	.62	93	.66	80	.65	82
15	.78	98	.46	83	.65	80	.90	68
16	.12	98	.36	94	.60	93	.57	91
17	.12	97	.46	93	.58	94	.75	91
18	.50	95	.18	91	.48	90	.44	86
19	....	100	.74	99	.59	97	.69	94
20	.26	86	.22	83	.28	87	.26	92
21	.45	76	.49	74	.39	61	.34	68
22	.44	95	.45	97	.65	87	.56	86
23	1.00	98	.27	97	.29	92	.53	85
24	.26	86	.48	85	.41	80	.40	77
25	.22	73	.32	73	.24	65	.20	70
26	.39	92	.10	91	.58	94	.56	90
27	....	100	.47	99	.78	99	.65	96
28	.06	96	.36	95	.32	91	.80	88
29	.52	98	—.01	97	.29	92	.34	95
30	.03	97	.23	99	.70	91	.63	89
31	.77	96	.54	98	.50	90	.77	86
32	.92	98	....	100	.64	97	.58	94
33	.45	86	.24	86	.48	82	.25	81
34	.19	96	.08	97	....	100	.90	98
35	.35	92	.30	88	.69	85	.60	71
36	....	100	.37	97	.49	94	.52	88
37	.58	73	.37	61	.62	62	.64	52
38	.52	87	.64	92	.72	74	.70	63
39	.29	77	.40	88	.62	76	.53	76
40	.52	49	.41	24	.12	24	.44	12
41	.19	91	—.02	95	.71	79	.83	76
42	.51	90	.06	91	.65	74	.47	74
43	.47	73	.37	76	.66	59	.48	44
44	.62	61	.46	52	.70	36	.60	34
45	.47	61	.38	47	.60	39	.68	27
46	.48	88	.41	83	.39	85	.53	77
47	.48	76	.38	68	.39	68	.51	78
48	.26	85	.19	62	.52	83	.48	60
49	.45	70	.34	70	.54	56	.44	53
50	.56	69	.49	79	.52	51	.62	44
51	.45	71	.18	72	.45	59	.52	63
52	.59	89	.13	82	.40	84	.56	80
53	.14	94	.22	93	.57	90	.47	85
54	.52	86	.62	76	.59	62	.75	61
55	.42	76	.48	69	.54	60	.27	65
56	.29	79	.61	52	.64	64	.65	45
57	.29	79	.62	91	.63	59	.72	64
58	.55	65	.71	57	.81	44	.72	40
59	.18	32	.11	29	.28	23	.15	17
60	.73	67	.60	62	.64	55	.82	44
61	.72	74	.59	45	.71	55	.81	36

TABLE II—Continued

VALIDITIES AND DIFFICULTIES OF THE ITEMS IN (HIGHER) FORM A OF THE OTIS SELF-ADMINISTERING TEST FOR A COLLEGE AND AN ADULT POPULATION[2]

Item No.	Crooks-Ferguson Men r	Crooks-Ferguson Men Per Cent Correct	Crooks-Ferguson Women r	Crooks-Ferguson Women Per Cent Correct	Chapanis Men r	Chapanis Men Per Cent Correct	Chapanis Women r	Chapanis Women Per Cent Correct
62	.81	70	.46	73	.83	59	.61	48
63	.59	78	.43	79	.77	68	.73	61
64	.45	74	.57	71	.71	54	.75	45
65	.79	89	.55	87	.81	64	.85	60
66	.62	67	.38	60	.74	46	.83	31
67	.61	44	.59	36	.82	28	.80	20
68	.66	41	.50	27	.50	25	.47	24
69	.73	58	.77	53	.67	43	.69	32
70	.42	28	.65	33	.60	20	.60	18
71	.76	67	.69	64	.68	54	.68	47
72	.62	43	.45	31	.94	34	.81	21
73	.76	44	.59	41	.87	28	.70	25
74	.67	44	.67	37	.66	34	.65	29
75	.60	30	.64	15	.78	16	.80	12

[2] The data for the college population are our own. Those for the adult population are from Chapanis.

10. There are (according to the writers' data) 30 items for college men and 22 items for college women which have validities of .50 or over.

We may conclude from the above facts that: (1) there is no difference in item difficulty between adults and college students; (2) there are no particularly striking sex differences in item difficulty either for college students or for adults; and (3) item validities tend to be somewhat lower for a college group than for an unselected group of adults.

A point of interest not brought out in the above discussion is the apparent existence of a tendency for item validity to vary inversely with item difficulty. Considering first the responses of men, the median difficulty for 30 items with validities of .50 or over is .72, while that for 18 items with non-significant validities is .95. The corresponding medians for women are in the first instance (based on 22 items) .62, and in the second instance (based on 23 items) .92. Since, however, the non-valid items on any test are often those which are passed by more than 90% or less than 10% of a group, the actual relation may not be inverse. It may appear so here only because there are so many easy items and so few hard ones.

In Table III are given results secured on Form B of the Otis Test. They appear to be quite similar to those already reported for Form A. The median sex difference is .05; the median item difficulty for men is .83 and that for women is .77; there are 21 items for college men and 14 items for college women which have non-significant validities and 22 items for college men and 27 items for college women which have validities of .50 or over (median difficulties are .34 and .47 for men and women respectively).

We agree with Chapanis that with respect to their validity certain of the items are misplaced, but because of inconsistencies between Chapanis', Otis' and our data we are unable to determine any one final correct order. It is interesting to note that the items (59, 25, 20, 29 and 6 of Form A) which Chapanis reports as the least valid items for the population he studied are also among the non-valid items of our study (for either men or women or both). The items (1, 34, 72, 65, and 67 of Form A) which he reports as most valid, however, are not among those which we have found occupying that position.

There are a number of items which appear in both forms A and B, in either identical wording or slightly modified wording but with the same essential content. For men the median difference in the item validities of 11 of these items having significant correlations is .09, and for women the median difference of 12 items is .11.

We shall not attempt to summarize the conclusions presented above, for that would involve mere repetition, but we should like to point out what we feel to be the most important of them. These are that the validity and difficulty of the items in the Otis test are lower for a college group than for an adult population.

TABLE III

VALIDITIES AND DIFFICULTIES OF THE ITEMS IN (HIGHER) FORM B OF THE OTIS SELF-ADMINISTERING TEST FOR A COLLEGE POPULATION

Item No.	Men r	Men Per Cent Correct	Women r	Women Per Cent Correct	Item No.	Men r	Men Per Cent Correct	Women r	Women Per Cent Correct
1	.16	99	....	100	44	.47	64	.51	42
2	.27	85	.45	74	45	.37	51	.34	36
3	.16	94	.38	91	46	.12	41	.43	38
4	.37	92	.74	79	47	.06	83	.15	84
5	.24	88	.48	77	48	.27	56	.30	55
6	.23	73	.32	82	49	.59	70	.65	48
7	.29	99	.04	94	50	.32	84	.72	90
8	.55	91	.61	77	51	.21	60	.17	58
9	.28	94	.23	80	52	.19	80	.17	73
10	....	100	.77	98	53	.67	68	.49	63
11	.69	93	.54	97	54	.21	75	.35	71
12	.12	92	.58	90	55	.69	84	.68	80
13	.40	83	.06	83	56	.39	59	.47	55
14	.24	96	.66	94	57	.39	70	.40	56
15	.10	69	.06	78	58	.59	59	.56	47
16	.34	96	.24	94	59	.46	72	.75	68
17	.18	89	.12	80	60	.61	84	.86	83
18	.22	84	.37	76	61	.88	88	.91	84
19	.53	89	.47	92	62	.68	74	.54	65
20	.36	91	.40	87	63	.22	47	.43	43
21	.14	93	.33	95	64	.67	53	.62	49
22	.21	88	.27	92	65	.71	54	.49	35
23	.12	83	.45	86	66	.52	45	.61	43
24	.42	91	.62	88	67	.76	50	.59	55
25	.33	90	.14	92	68	.75	76	.53	14
26	.20	94	.49	95	69	.67	38	.53	43
27	.40	80	.45	71	70	.70	45	.66	53
28	.34	96	.16	94	71	.65	25	.65	37
29	.17	83	.43	74	72	.67	41	.69	62
30	.31	79	.35	58	73	.73	27	.66	45
31	.48	93	.63	65	74	.59	15	.43	22
32	.25	69	.42	80	75	.39	2	.15	1
33	.63	92	.21	96					
34	.40	94	.05	90					
35	.27	64	.48	60					
36	.23	85	.48	82					
37	—.05	80	.23	83					
38	.24	71	.28	75					
39	.39	73	.50	62					
40	—.25	97	.32	91					
41	.07	80	.50	75					
42	.15	69	.67	83					
43	.32	69	.43	59					

THE PEABODY LIBRARY INFORMATION TEST: A STUDY OF ITS STATISTICAL VALIDITY AND RELIABILITY

GEORGE H. DEER
Louisiana State University

Colleges and public schools alike have been criticized for failure to teach their pupils an efficient and intelligent mastery in the use of the library. One probable factor in this tendency to neglect so important a function in the pupil's growth and development has been the lack of an adequate measuring device for the skills and information involved. Until recently no satisfactory comprehensive test had been devised for use in the teaching of library information or library use.

The Peabody Library Information Test, devised by Dr. Louis Shores, head of the library school, and Dr. Joseph E. Moore, professor of educational psychology, George Peabody College for Teachers, contains 188 multiple choice, completion, and matching items. In its revised form the test consists of eight sections, each of which is really a separate, independent test. These consist of (1) The Book, 24 items; (2) Arrangement of Books, 24; (3) The Card Catalogue, 24; (4) The Dictionary, 27; (5) The Encyclopedia, 24; (6) Periodicals and Indexes, 20; (7) Special Reference Books, 24; and (8) Bibliography, 23.

Much as such a test is needed in this field, none can be accepted as desirable simply because it happens to be available. Careful users of a test will insist that its worth must be established by its validity and its reliability. This study is an attempt to examine the validity and the reliability of the Peabody Library Information Test when administered to 1322 students in nine colleges in situations involving actual classroom use. After all, the user of a test is not concerned with what might be the highest coefficient of reliability the author of a test has been able to establish under ideal conditions, but he is interested in the degree of reliability which may usually be expected for the test under everyday classroom conditions.

The 1322 subjects for this test were obtained from the freshman classes of nine colleges, as follows: Bethel College, 117; David Lipscomb College, 143; Fisk University, 111; Furman University, 241; Ouachita College, 115; Peabody College, 121; Stetson University, 198; Tennessee Polytechnic Institute, 81; Western Kentucky Teachers College, 135; and 59 juniors from Peabody College. The tests were administered in the regular English classes.

Table 1 shows the number of students, measures of central tendency, and measures of variability for the Freshmen of each college.

VALIDITY

This study is not concerned with the procedures used by the authors while constructing their instrument in order to insure its validity; it is concerned only with the statistical measures which indicate validity or non-validity for the test when used in the classroom.

TABLE 1

A COMPARISON OF THE MEASURES OF CENTRAL TENDENCY, VARIABILITY, AND RANGE FOR THE FRESHMAN CLASSES OF NINE COLLEGES FOR SCORES MADE ON THE PEABODY LIBRARY INFORMATION TEST

College	A	B	C	D	E	F	G	H	I
Number Subjects	117	143	111	241	115	121	198	81	135
Median	86.3	62.8	61.1	58.4	50.2	31.0	63.8	54.8	76.1
Q_1	74.0	50.3	51.4	46.5	41.4	21.4	52.6	47.0	62.3
Q_3	100.0	75.0	74.0	72.8	61.3	42.5	76.4	69.0	95.0
Q	14.5	12.4	11.3	13.2	9.9	10.5	11.9	11.0	16.4
Mean	85.0	64.4	63.5	60.5	51.9	32.3	66.0	58.3	74.8
σ	17.5	18.6	17.0	17.8	15.5	15.8	18.5	16.6	21.5
σM	1.6	1.5	1.6	1.1	1.5	1.5	1.3	1.8	1.9
Range	43–123	8–113	18–98	23–123	13–88	5–62	13–123	23–103	28–128

TABLE 2

DIFFERENCES BETWEEN THE MEANS FOR SCORES MADE BY FRESHMAN CLASSES OF NINE COLLEGES ON THE PEABODY LIBRARY INFORMATION TEST. ALSO CRITICAL RATIOS FOR THESE DIFFERENCES

College	B		C		D		E		F		G		H		I	
	D	CR	D	CR	D	CR	D	CR	D	CR	D	CR	D	CR	D	CR
A	20.6	9.4	22.5	9.8	24.2	12.0	33.1	15.0	52.7	24.0	19.0	9.5	26.7	11.1	10.2	4.3
B			1.9	.9	3.9	2.0	12.5	5.9	32.1	15.3	1.6	.8	6.1	2.5	10.4	4.3
C					2.0	1.0	10.6	4.8	30.2	13.8	3.5	1.7	4.2	1.7	12.3	4.9
D							8.6	4.5	28.2	13.5	5.5	4.2	2.2	1.0	14.3	6.5
E									19.2	9.4	14.1	7.1	6.4	2.7	22.9	9.2
F											33.7	16.8	26.0	10.7	42.5	17.7
G													7.7	3.5	8.8	3.8
H															16.5	6.3

The first measure of validity applied to the test here is to determine whether it really discriminates between the different groups to which it is administered. Table 2 presents the differences between the means for each of the nine colleges and the mean of every other college. Differences were found to be as high as 52.7 between the mean for school A and that for school F. The column of critical ratios is most significant in this table, showing that 27 of the 36 differences meet the general criterion of reliability, 3 sigma. These critical ratios run as high as to 24 for differences of means for schools A and F. This indicates that the test makes a high degree of discrimination between groups in the different colleges.

A further test of discrimination was applied by comparing scores made by the freshman group in College F with those made by 59 juniors in the same school who had no formal library instruction in order to determine whether the test would discriminate between freshmen and more advanced classes. One would naturally expect that in two years of study the college student would acquire incidentally a considerable amount of information about the library and its use. A test sensitive to different degrees of knowledge should be able to recognize this additional amount of library information acquired between the freshman and the junior year. Results of this comparison are presented in Table 3, which shows a difference of 15.6 in the means for the two groups, with a critical ratio of 6. This indicates a rather distinct differentiation between classes at different college levels.

As a further check on the ability of the test to make discriminations, these scores for the college juniors were compared with the scores made by the same individuals after a two weeks' period of instruction in the use of the library, presented in the regular junior English classes. This two weeks of instruction increased the group's mean from 47.9 to 97.5, a difference of 49.6 with a critical ratio of 20.2. This would seem to point to the fact that the test was measuring, at least, whatever the instructor was teaching as library usage. These data are presented in Table 4.

TABLE 3

MEANS, SIGMAS AND CRITICAL RATIOS FOR FRESHMEN AND JUNIORS IN COLLEGE F ON THE PEABODY LIBRARY INFORMATION TEST

	Freshmen		Juniors
Number	121		59
Mean	32.3		47.9
σ	15.8		16.6
σ_M	1.5		2.2
Difference between Means		15.6	
Standard Error of Difference		2.6	
Critical Ratio		6.0	

TABLE 4

COMPARISON OF SCORES MADE BY COLLEGE JUNIORS WITH SCORES MADE BY THE SAME INDIVIDUALS AFTER TWO WEEKS OF LIBRARY INSTRUCTION

	Initial Test		Retest After Two Weeks Instruction
Number of Subjects	59		59
Mean Score	47.9		97.5
Standard Deviations	16.6		25.5
Standard Error of the Mean	2.16		3.3
Difference between Means		49.6	
Correlation, initial and final test		.68	
Standard Error of D_{M1-M2}		2.45	
Critical Ratio		20.2	

TABLE 5

PERCENT OF FRESHMEN IN THE SCHOOLS LISTED HORIZONTALLY WHO EQUAL OR EXCEED THE MEDIAN SCORE OF THE FRESHMEN IN EACH SCHOOL LISTED VERTICALLY

College	A	B	C	D	E	F	G	H	I
A		11	9	8	2	0	14	7	31
B	87		46	41	22	8	52	27	71
C	89	54		44	25	9	56	38	73
D	91	59	59		31	81	62	43	75
E	98	71	80	68		15	81	63	83
F	100	97	98	96	90		99	97	99
G	87	48	43	40	21	6		27	70
H	97	73	69	59	40	10	74		80
I	72	24	21	18	9	2	79	15	

TABLE 6

INTERCORRELATIONS OF THE EIGHT PARTS OF THE PEABODY LIBRARY INFORMATION TEST ON SCORES MADE BY 117 FRESHMEN IN COLLEGE A*

Parts	II	III	IV	V	VI	VII	VIII
I	.53	.42	.33	.24	.39	.26	.31
II		.55	.36	.26	.36	.11	.35
III			.19	.15	.25	.23	.27
IV				.27	.44	.22	.34
V					.34	.10	.10
VI						.32	.20
VII							.27

* No coefficient has a probable error smaller than .05 nor larger than .06.

A commonly used measure of the degree of discrimination a test possesses is the percent of overlapping from one group to another. A fifty percent overlapping for one group over the median of another would indicate that the test was not measuring a difference between the groups.

Table 5 shows an overlapping ranging from zero to one hundred percent between the different colleges. Not a single student in one college ranked as high as the median of a certain other college and not a student in this second school scored as low as the median in the first.

In a test made up of several parts, as this one is, it is important to know whether the separate parts of the test are really independent and measuring different things; otherwise its great length cannot be justified. To determine this, intercorrelations were computed for all of the parts, using the students of College A as subjects. This group was chosen rather than another because of their much higher total scores on the test. The coefficients in Table 6 denote relatively little relation between the different sections of the test. Each of the sections is measuring a different function which cannot be measured by any one of the other parts.

RELIABILITY

As a measure of the reliability of the Peabody Test a comparison was made of scores obtained on alternate items. This was done by correlating the scores made on the even items with those made on the odd items and then applying the Spearman–Brown correction formula. Corrected coefficients were obtained as follows: Peabody Freshmen, .68; Peabody Juniors, .82; Peabody Juniors after two-weeks teaching, .96; T.P.I., .81; Ouachita College, .92; Bethel College, .88.

The probable error of estimate of a raw score was found to range from 1.75 to 4.5.

SUMMARY

This study is an analysis of the validity and reliability of the Peabody Library Information Test as administered to 1263 freshmen and 59 juniors under ordinary classroom conditions in nine colleges, when accepted statistical measures of validity and of reliability are applied to it.

The findings indicate that the test is capable of a high degree of discrimination between freshmen classes of different schools as well as between classes at different scholastic levels within the same college. It also

measures a distinct difference in the amount of knowledge possessed by a junior class before and after two weeks' instruction in library usage.

The eight separate sections of the test show such low intercorrelations as to indicate that each is measuring a separate function which is not measured by either of the other seven parts.

In only one group was the reliability coefficient found to be below .81, as obtained by correlating the even items with the odd by the Pearson product moment method and applying the Spearman–Brown correction formula. This coefficient of reliability was found to be as high as .96 for one group.

The P. $E._{est}$ for an individual's raw score ranged from 1.75 to 4.5.

AN ANALYSIS OF THE RESULTS FROM THE FIRST ANNUAL EDITION OF THE NATIONAL TEACHER EXAMINATIONS

JOHN C. FLANAGAN
Cooperative Test Service

INTRODUCTION

This discussion is a report of the results of the first annual administration of the National Teacher Examinations. Discussions of the nature of the examination service and its development[1] and of the administration of these examinations[2] have been published elsewhere. The National Committee on Teacher Examinations was appointed by the American Council on Education at the request of a group of superintendents in the spring of 1939. The Carnegie Foundation for the Advancement of Teaching provided funds for the development of the project, and the National Committee assigned to the Cooperative Test Service of the American Council on Education the task of preparing the annual forms of the battery of examinations to be used.

The service made available to superintendents and boards of education by the National Committee on Teacher Examinations has three principal objectives: first, the provision and administration of a battery of teacher examinations; second, the collation and distribution of information on the appraisal of other factors to be considered in the selection of teachers; third, the promotion of research in teacher education and guidance.

It should be stated at the outset that the superintendents who inaugurated this project were well aware that no battery of tests could at the present time make possible a complete evaluation of all factors to be included in the selection of teachers. However, this group of superintendents felt that having objective and comparable information on certain aspects of training and professional preparation would be of great assistance to them with their individual problems of teacher selection. In common with all other persons who have been confronted with the task of teacher selection, these superintendents fully realized the difficulty of interpreting the types of information which have previously been available to them.

The selection of general areas to be tested was made chiefly through a series of conferences with superintendents and a questionnaire sent to a larger group of superintendents in the spring of 1939. Information about such traits as desirable attitudes and personal characteristics had to be eliminated from the proposed battery of examinations since satisfactory techniques for measuring these traits objectively are not available. In general, the superintendents stated that they wished information about the extent to which candidates could read with understanding, could express themselves clearly, and could handle numerical and non-verbal concepts intelligently. They also wished to know to what extent the candidates had acquired a good liberal education and the extent of their awareness of contemporary affairs and current social problems. They wished information concerning the degree to which the candidate was aware of the social implications of education, his knowledge of the basic principles of psychology as they apply to education, and his familiarity with the techniques of guiding and assisting students. Lastly, they wished to know the extent to which the teacher was really a master of those fields in which he proposed to teach.

Starting with this general outline, for which these superintendents also supplied approximate time allotments, the staff of the Cooperative Test Service began in the summer of 1939 the preparation of the first battery of examinations.

Over a period of ten years the Cooperative Test Service has developed a method of test construction which has proved quite effective. Following this method, each step in the construction of the tests is performed by those individuals most competent for this particular task. The method utilizes four groups of individuals: first, editors in the various fields in which examinations are being constructed who are experts in the subject matter of those fields and who have unusual

[1] *An Announcement of a Teacher Examination Service*, National Committee on Teacher Examinations of the American Council on Education, 15 Amsterdam Avenue, New York City, November 1939.

[2] Ryans, David G. "The Professional Examination of Teaching Candidates: A Report of the First Annual Administration of the National Teacher Examinations," *School and Society*, LII (October 5, 1940), pp. 273–284.

ability to express ideas clearly and concisely; second, scholars, supervisors, and other experts and authorities who are selected from the entire country and who are responsible for the planning and final editing of the various examinations; third, experts in the techniques of test construction who assist the editors and committees with the technical aspects of the preparation of the examinations; fourth, a large group of test users who act as test critics and collaborators and who insure that the final form of the examination is practical as well as valid.

The steps involved in the actual construction of these examinations may be briefly outlined as follows:

1. Developing outlines and specifications for the various examinations to be constructed, in cooperation with a group of authorities and experts.
2. Writing test items in accordance with the specifications and outlines, and editing the items.
3. Administering these preliminary items to students in teacher training institutions to eliminate weak and ambiguous items or items that are too easy or too difficult.
4. Selecting and revising the items for the tentative final forms of the examinations.
5. Obtaining criticisms and editorial assistance from a large number of supervisors, heads of departments, scholars, experts in teacher training, and similar individuals.
6. Editing on the basis of criticisms received.

As mentioned in the outline above the initial work on the development of the outline specifications for each of the various tests was done by the staff editor in the particular field. This tentative outline was prepared after several weeks of careful research into what experts and school officials consider should be the relative attention given to the various sub-topics which are included under each of the major classifications into which teacher education is divided. Courses of study, textbooks, journal articles, and reports by various professional organizations were analyzed carefully, and these findings were briefly summarized for the convenience of the experts and authorities who were cooperating in the preparation of the outline for the particular test. These materials were sent out to about 150 individuals, who were asked to suggest changes in the tentative outlines or in the specifications tentatively drawn up for the writing of the examination questions. An effort was made to have this group of cooperating individuals as representative as possible of all the interested groups. Because of the shortness of time allowed for the preparation of the first form, it was impossible to make as thorough a canvass of interested groups as it is expected will be made with later forms, but it was considered that a fairly representative group was obtained. On the average there were about eleven people who cooperated in the preparation of each of the outlines. The total group was divided as follows:

Teachers' college officials	18 per cent
Teacher training personnel in other colleges and universities	21 per cent
Instructors in academic fields not principally concerned with teacher training	18 per cent
Teachers and supervisors in city public school systems	42 per cent

In writing the items for the various examinations, an effort was made to test for the ability of the candidate to *apply* to practical school situations, or, in some cases, to new situations of other varieties, the facts and principles which he had learned. In doing this, much use was made of the best answer technique; that is, the situation was described and several possible answers, usually five, were provided. The candidate was asked to select the answer which he regarded as best. The preparation of such items is, of course, very much more difficult than the writing of items which test the simple recognition of facts as they have been taught; but the resultant test is generally regarded as a much more valid indication of the teacher's ability to show good judgment in an actual teaching situation. The inclusion of only those items upon which the experts are in agreement concerning the best answer appears to be the most satisfactory method available at present of providing a test which will be fair to the applicants and at the same time enable the employing officials to discriminate among candidates on this basis.

The examinations prepared in this way were administered to 3726 candidates in twenty-three examining centers on March 29–30, 1940. The examination centers participating in this first administration of the National

Teacher examinations were carefully selected on the basis of the ability of the school systems to take full advantage of the examination service. A number of the cities had previously been using examinations as part of their teacher selection program, and merely substituted the National Teacher Examinations for the tests previously used. After administration the examinations were all returned to the central office for scoring by means of the International Test Scoring machine. The scores were reported to the school systems and to the candidates on April 30. A general report of the administration of the examinations has been prepared by Dr. David G. Ryans.[8]

Results

The most obvious finding from the administration of these examinations is that *candidates for teaching positions are not equally well qualified.* The distributions of scores are shown in Table I. Inspection of the table makes it clear that although all of these individuals have satisfied the legal requirements for certification and have accumulated the necessary credits, some of them know so little in various fields that they are able to answer correctly practically none of the questions in the examinations, whereas others make very nearly perfect scores on the examinations.

Although a very large amount of evidence has accumulated through other testing programs, such as the National College Sophomore Testing Program sponsored by the American Council on Education and the testing program of the Teachers College Personnel Association, showing that large differences exist between institutions, it seems worth while to add evidence from this particular testing program. The distributions of total scores on the common examinations for seven institutions from each of which there were more than seventy-five students who took the National Teacher Examinations are shown graphically in Figure 1. The chart shows the scores achieved by the students from two state universities, three private universities, and the teachers' colleges of two different states. It should be emphasized that these charts are based on only those students from the various institutions who took the teacher examinations this past spring. Although these groups are therefore not known to be representative of the entire student body or even the group of students prepared to be teachers by the particular institution, the charts show that of the students presenting themselves for examination eighty-five per cent of those from one institution did better than did the average student in two of the other institutions. It is quite likely that some of the differences would be even larger were records on the entire graduating groups available for all these colleges. However, it should not be overlooked that these results indicate that there are some very good studies in all of these colleges.

A second point of considerable interest which is brought out by an analysis of the results is that *a fairly long and varied battery of examinations is necessary to describe the various candidates adequately.* The first evidence which will be presented on this point is the profiles on the battery of examinations shown in Figure 2. This figure is reproduced from the profile chart which was sent to each individual who took the examinations with the suggestion that he plot his own profile on the chart in order to interpret more readily his own pattern of achievement. The ten profiles on the chart, representing the average scores made on the various common examinations by the group of candidates taking a particular subject examination, clearly emphasize the marked differences in pattern of these groups, even in cases where the combined averages are very similar. The profiles are self-explanatory and seem to be about what one would expect. The Mathematics and Physical Science groups are outstanding on the Reasoning Test, the Language groups are superior on the English Expression Test, and the Social Studies, Science, and Mathematics groups are high on their respective sections in the General Culture Test.

Although all the various examinations in the battery are positively correlated, the fact that each one is making a unique contribution to the total battery is shown by the intercorrelations of these various parts presented in Table II. The smallest of the Pearson product-moment correlation coefficients shown in the table is .23 between the Fine Arts and the Mathematics sections of the General Culture examination. The largest coefficient is .81 between the Current Social Problems and the History and Social Studies sections of that examination. The median of the fifty-five correlation coefficients in the table is .56. Pairs of tests which would be expected to show substantial correlation, such as Fine Arts and

[8] *Ibid.*

TABLE I

DISTRIBUTIONS OF THE SCORES OF THE CANDIDATES TAKING THE NATIONAL TEACHER EXAMINATIONS IN MARCH, 1940, IN TERMS OF SCALED SCORES AND RAW SCORES

*Scaled Score							General Culture												Prof. Inf.			
Max. Possible Score	Reasoning *Raw Score	No.	Eng. Comp. Raw Score	No.	Eng. Exp. Raw Score	No.	Curr. Soc. Prob. Raw Score	No.	History Raw Score	No.	Literature Raw Score	No.	Science Raw Score	No.	Fine Arts Raw Score	No.	Math. Raw Score	No.	Elem. Raw Score	No.	Secondary Raw Score	No.
	100		*120*		*157*		*75*		*75*		*90*		*60*		*90*		*60*		*240*		*240*	
100	*87*	2			*147*	1															*204*	1
98	*85*	1			*145*	0															*199*	3
96	*83*	0			*142*	3	*72*	1			*88*	3	*56*	2	*87*	3					*195*	6
94	*81*	6	*112*	1	*140*	7	*71*	1	*70*	6	*86*	4	*55*	1	*85*	5			*194*	1	*189*	6
92	*79*	4	*109*	10	*137*	13	*70*	4	*68*	16	*84*	10	*53*	12	*83*	8	*60*	14	*189*	1	*184*	8
90	*77*	15	*106*	14	*134*	17	*68*	25	*66*	16	*82*	20	*51*	9	*81*	9		0	*184*	0		
88	*75*	14	*102*	39	*131*	42	*67*	3	*65*	17	*79*	31	*49*	26	*79*	26	*59*	14	*179*	4	*180*	8
86	*73*	21	*99*	23	*128*	26	*65*	34	*63*	34	*77*	26	*47*	38	*77*	31	*57*	45	*174*	1	*175*	23
84	*71*	29	*96*	54	*126*	34	*64*	29	*61*	37	*74*	66	*45*	36	*75*	46	*56*	28	*169*	4	*170*	38
82	*69*	39	*93*	67	*123*	61	*62*	68	*59*	78	*71*	64	*43*	59	*72*	84	*54*	70	*164*	5	*165*	54
80	*67*	61	*89*	125	*120*	85	*60*	93	*57*	75	*68*	78	*41*	53	*70*	65	*51*	67	*159*	7	*159*	98
78	*65*	92	*86*	114	*117*	105	*58*	127	*55*	77	*65*	90	*39*	66	*67*	149	*47*	104	*154*	14	*154*	110
76	*63*	94	*82*	167	*114*	107	*56*	106	*52*	154	*62*	123	*37*	95	*65*	111	*43*	119	*149*	13	*149*	109
74	*61*	104	*79*	145	*111*	147	*54*	172	*50*	142	*59*	154	*35*	97	*62*	171	*39*	142	*144*	21	*144*	134
72	*59*	103	*75*	220	*109*	100	*52*	161	*48*	187	*56*	160	*33*	148	*60*	129	*36*	125	*139*	19	*139*	137
70	*56*	193	*71*	233	*106*	181	*50*	155	*46*	131	*52*	257	*31*	121	*57*	219	*33*	159	*134*	27	*134*	152
68	*54*	151	*68*	191	*103*	199	*48*	200	*43*	260	*49*	216	*29*	179	*54*	245	*30*	185	*129*	32	*129*	167
66	*52*	179	*64*	255	*100*	214	*46*	198	*41*	181	*45*	276	*27*	209	*51*	250	*27*	231	*124*	33	*124*	196
64	*50*	197	*61*	227	*97*	254	*43*	306	*39*	220	*42*	205	*25*	216	*49*	193	*25*	172	*119*	52	*119*	206
62	*47*	315	*57*	257	*94*	211	*41*	204	*36*	305	*39*	229	*23*	241	*46*	236	*22*	304	*114*	47	*114*	173
60	*45*	211	*53*	226	*91*	221	*39*	201	*34*	225	*36*	222	*21*	303	*43*	259	*20*	230	*109*	36	*108*	223
58	*42*	305	*49*	273	*88*	226	*36*	290	*32*	194	*33*	30	*20*	172	*40*	243	*18*	250	*104*	35	*103*	175
56	*40*	187	*46*	164	*84*	229	*34*	209	*29*	258	*30*	233	*18*	284	*37*	256	*16*	233	*98*	48	*98*	145
54	*37*	298	*42*	208	*81*	156	*32*	150	*27*	176	*27*	203	*16*	266	*35*	129	*14*	247	*93*	38	*93*	134
52	*34*	218	*39*	138	*78*	147	*29*	224	*25*	162	*24*	200	*15*	154	*32*	176	*12*	216	*87*	44	*87*	139
50	*31*	162	*35*	156	*74*	199	*27*	114	*23*	160	*22*	101	*13*	232	*29*	165	*11*	106	*81*	42	*81*	134
48	*28*	169	*32*	89	*70*	160	*25*	123	*21*	118	*20*	92	*12*	102	*26*	124	*9*	182	*75*	37	*75*	97
46	*26*	76	*29*	92	*66*	132	*23*	116	*19*	107	*17*	135	*11*	103	*24*	68	*8*	70	*70*	22	*69*	78
44	*23*	107	*25*	62	*62*	113	*21*	87	*17*	74	*15*	55	*9*	168	*21*	76	*6*	128	*64*	31	*63*	58
42	*20*	79	*22*	29	*58*	84	*19*	60	*15*	67	*13*	53	*8*	56	*19*	48	*5*	46	*58*	18	*57*	57
40	*18*	54	*19*	30	*53*	73	*16*	67	*13*	42	*11*	24	*7*	45	*16*	38	*4*	35	*52*	12	*50*	40
38	*15*	61	*16*	7	*49*	32	*14*	28	*11*	35	*10*	10	*6*	40	*14*	14	*3*	31	*45*	10	*43*	26
36	*13*	36	*13*	8	*45*	16	*12*	23	*9*	22	*8*	16	*5*	25	*11*	13	*2*	15	*39*	2	*37*	10
34	*10*	25	*11*	4	*41*	10	*11*	6	*8*	5	*7*	6	*4*	16	*9*	5		0	*32*	3	*30*	15
32	*8*	10	*8*	4	*38*	7	*9*	8	*6*	10	*5*	4	*3*	11	*7*	3	*1*	16	*25*	1	*23*	5
30	*6*	11	*5*	1	*34*	8	*7*	5	*4*	8	*4*	2	*2*	5	*5*	2	*0*	17	*18*	1	*16*	0
28	*4*	1	*3*	0	*30*	3	*5*	3	*3*	1	*3*	1	*1*	6	*3*	1			*9*	1	*8*	1
26	*1*	3	*1*	1	*26*	8	*4*	0	*1*	1	*2*	1	*0*	6	*1*	2						
24	*0*	2	*0*	1	*23*	2	*2*	1	*0*	1	*1*	1										
22					*19*	0					*0*	1										
20					*16*	2																
No.	3635		3635		3635		3602		3602		3602		3602		3602		3601		662		2966	
Mdn.	44.63		61.39		93.36		41.34		36.83		41.26		21.23		47.77		21.31		106.50		117.72	
M	44.13		61.37		92.20		41.02		37.09		42.38		22.65		47.66		24.18		105.98		116.25	
σ	14.29		19.67		19.89		12.68		12.99		16.94		10.34		15.13		13.85		29.68		31.02	

* The value given is the lowest value falling within the particular interval.

TABLE I

DISTRIBUTIONS OF THE SCORES OF THE CANDIDATES TAKING THE NATIONAL TEACHER EXAMINATIONS IN MARCH, 1940, IN TERMS

M	44.13	61.37	92.20	41.02	37.09	42.38	22.65	47.66	21.31	106.50	117.72
σ	14.29	19.67	19.89	12.68	12.99	16.94	10.34	15.13	13.85	29.68	31.02

* The value given is the lowest value falling within the particular interval.

TABLE I

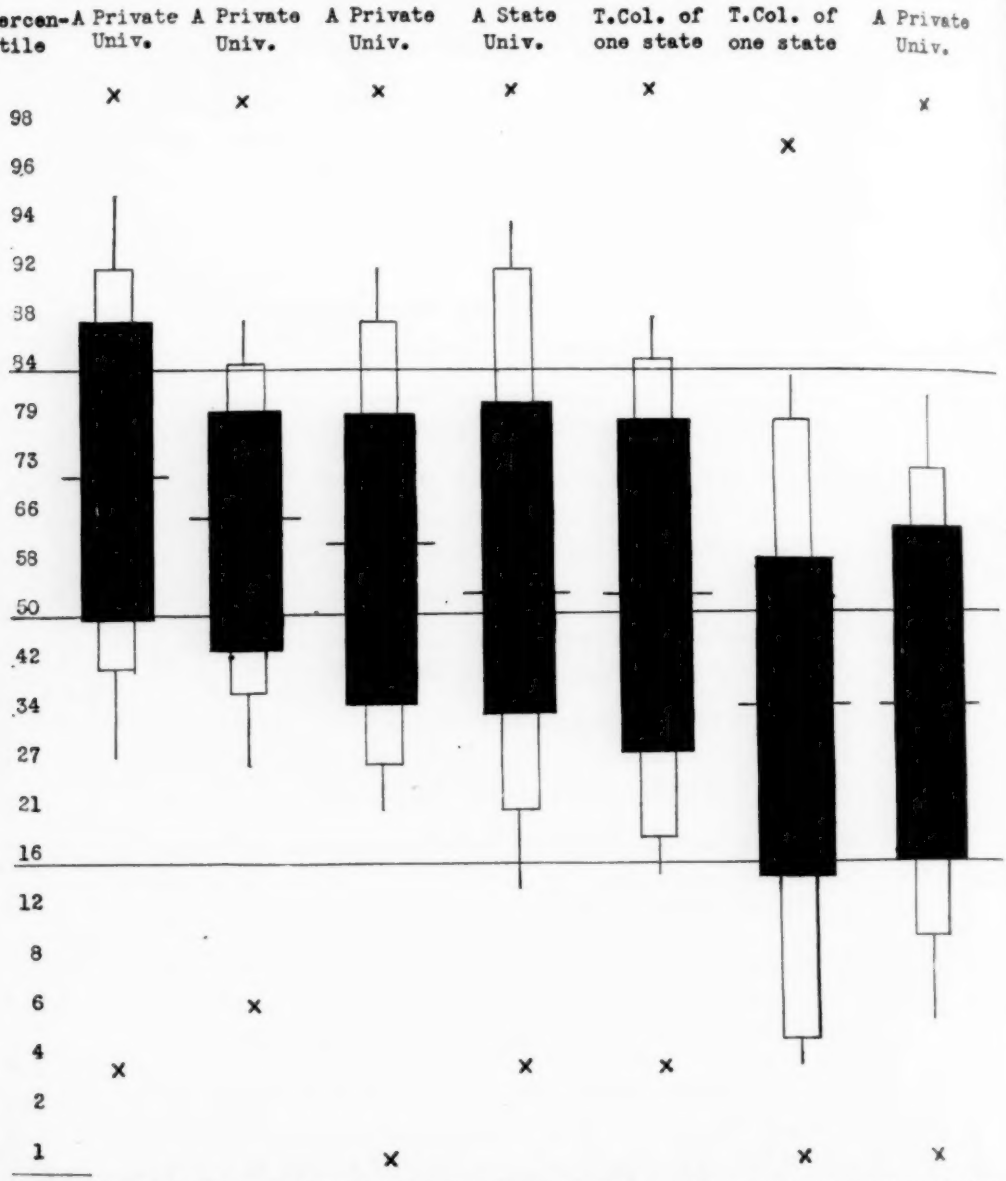
DISTRIBUTIONS OF THE SCORES OF THE CANDIDATES TAKING THE NATIONAL TEACHER EXAMINATIONS IN MARCH, 1940, IN TERMS OF SCALED SCORES AND RAW SCORES—Continued

*Scaled Score	Optional Examinations																					
Max. Possible Score	Cont. Affairs		Educ. in Elem.		Eng. Lang. & Lit.		Soc. Studies		Math.		Biol. Science		Phys. Science		French		German		Spanish		Latin	
	*Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.	Raw Score	No.
100	150		241		290		252		105		177		119		281		283		283		237	
98					251	2																
96	126	1			241	0			95	1											211	1
94	123	6			232	6			91	3			83	3	254	1					203	0
92	119	8	162	1	222	5	215	1	86	6			78	1	248	2					194	3
90	116	19	158	1	212	12	206	6	82	4			73	3	242	3			256	1	185	4
88	112	31	153	0	203	14	197	17	77	14	145	4	68	2	236	5	246	1	244	1	176	1
86	109	19	149	2	194	18	189	11	72	11	139	5	64	5	229	9	238	3	233	0	167	6
84	105	43	145	4	184	23	181	17	67	10	134	7	60	6	223	5	231	2	221	2	157	6
82	101	64	141	4	174	33	172	22	62	19	128	4	56	9	215	10	220	5	210	0	148	8
80	97	85	136	9	164	47	163	28	57	13	122	12	52	13	205	13	202	2	197	1	138	7
78	93	94	132	8	155	44	155	48	52	25	116	15	48	13	191	12	182	5	186	3	129	11
76	89	108	127	13	145	49	146	41	46	31	110	10	44	20	172	21	164	6	174	4	121	12
74	85	138	123	12	136	58	138	46	41	32	104	10	41	21	160	16	151	5	161	4	112	15
72	80	178	119	22	126	81	130	58	37	25	99	21	38	17	150	14	140	4	148	2	105	6
70	75	204	115	16	117	66	121	58	34	22	93	22	34	14	139	20	130	4	136	6	98	10
68	71	166	110	28	108	74	113	70	31	28	87	23	31	14	129	22	121	3	125	7	92	7
66	66	231	106	36	99	86	104	65	28	36	81	17	27	24	119	21	112	1	115	3	86	9
64	61	258	102	31	90	71	96	61	26	24	76	19	25	16	109	16	104	2	106	5	80	1
62	56	248	97	54	82	71	88	64	25	31	70	28	22	24	100	12	95	3	97	5	75	12
60	51	258	93	30	74	62	80	75	21	19	65	21	19	22	90	19	86	7	89	2	69	8
58	47	213	88	45	67	54	73	65	19	26	60	25	17	13	82	7	78	4	81	4	64	4
56	42	217	83	47	61	33	66	61	17	20	54	21	14	25	73	9	69	4	73	5	59	2
54	38	186	78	31	55	61	60	54	15	15	49	18	12	17	64	10	62	2	66	4	54	6
52	34	146	74	28	50	29	55	44	14	11	44	12	10	20	55	9	54	1	59	3	50	1
50	30	141	68	51	44	36	49	41	12	11	39	19	7	16	46	4	47	3	51	1	45	1
48	27	63	64	23	39	28	45	33	10	9	34	13	5	11	38	3	39	2	44	3	40	4
46	24	82	59	20	34	17	41	31	9	1	29	6	4	1	31	1	32	2	38	0	36	0
44	21	69	54	13	29	18	37	17	7	6	25	2	2	5	25	2	26	1	32	0	31	1
42	18	32	50	15	25	6	33	8	5	6	21	9	0	7	20	1	20	1	26	0	27	1
40	15	38	45	9	21	7	29	14	4	1	18	3			14	2	14	1	21	0	22	1
38	13	18	40	12	17	5	25	17	2	1	14	3			9	0			16	0		
36	11	12	36	9	14	2	21	8	1	1	11	1			5	0			12	1		
34	9	14	32	2	11	2	18	4	0	1	9	0			2	0						
32	7	4	28	1	8	1	15	3			6	1			0	1						
30	5	1	24	1	5	4	12	1														
28	4	1	20	1	3	1	9	1														
26																						
24																						
22																						
20																						
No.	3396		579		1126		1090		463		351		342		270		74		67		148	
Mdn.	59.60		90.45		104.26		96.20		31.84		74.11		25.75		136.68		129.50		116.17		111.50	
M	60.72		89.84		107.76		100.41		36.91		76.55		29.20		141.84		130.99		122.13		110.93	
σ	24.16		25.10		48.22		44.31		20.43		30.99		18.00		55.69		64.11		51.89		40.33	

* The value given is the lowest value falling within the particular interval.

FIGURE 1

THE GRAPHIC DISTRIBUTIONS OF THE TOTAL SCORE ON ALL THE COMMON EXAMINATIONS IN THE NATIONAL TEACHER EXAMINATIONS ACHIEVED BY THE CANDIDATES TAKING THE EXAMINATIONS IN MARCH, 1940, WHO HAD RECEIVED THEIR TRAINING AT CERTAIN SELECTED INSTITUTIONS

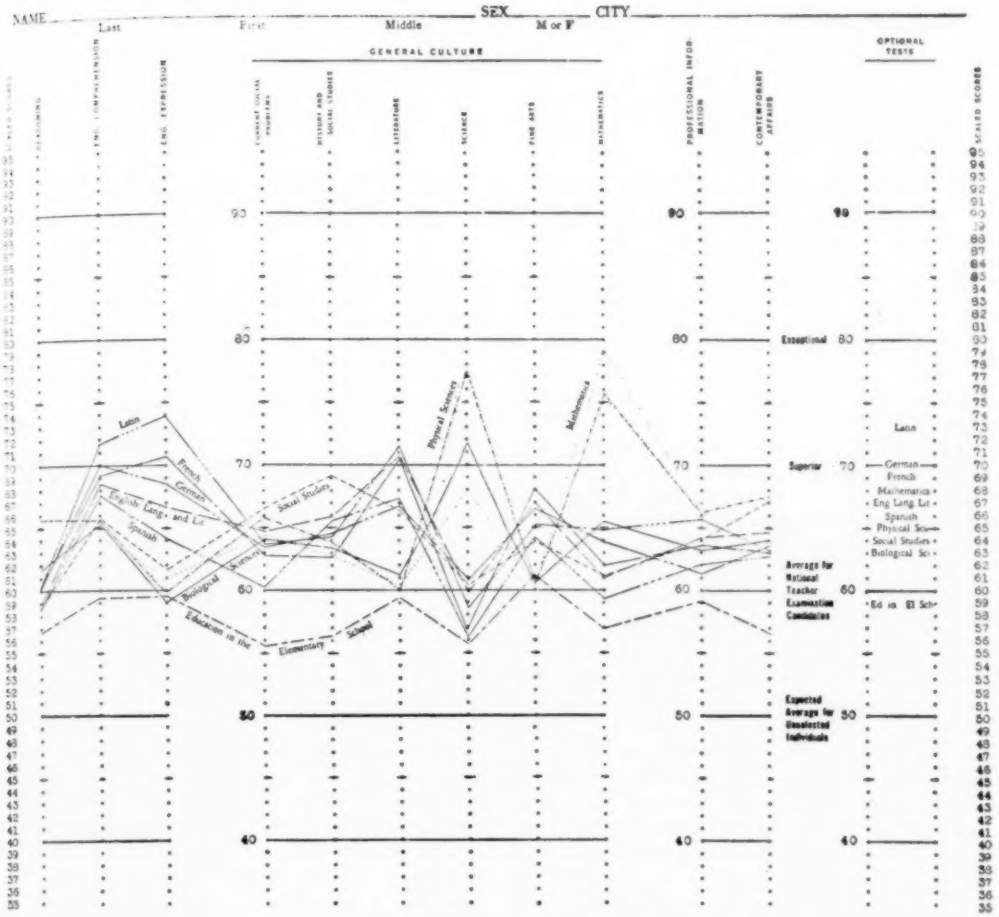


The percentile scale at the left is based on the returns for all the candidates taking the National Teacher Examinations. The middle horizontal lines show the national median, and the other two are at the 16th and 84th percentiles of the distribution. The wide portion of each bar represents the range of scores of the middle half of the candidates from the particular institution. The narrow parts extend to the 16th and 84th percentiles. The lines at the end extend down to the 10th percentile and up to the 90th percentile. The crosses represent the minimum and maximum scores for each institution. The short horizontal line near the middle of each bar represents the median score for the particular institution.

FIGURE 2

INDIVIDUAL PROFILE CHART FOR NATIONAL TEACHER EXAMINATIONS

May, 1940



DIRECTIONS. It is suggested that each candidate fill in his own test results as reported on the card sent to him May first. This may be done by drawing circles around the dots corresponding to his scores on the various subtests of the National Teacher Examination Battery.

Each of the ten profiles already drawn on the chart represents the average scores made on the various common examinations by the group of candidates taking a particular special subject examination. The code for identifying these special subject group profiles is shown below:

1. Education in the Elementary School
2. English Language and Literature
3. Social Studies
4. Mathematics
5. Biological Sciences
6. Physical Sciences
7. French
8. German
9. Spanish
10. Latin

VALUE OF THE CHART. This chart enables the candidate to appraise his own performances in various ways:

1. He can compare his score on any test with the Scaled Score 50, which is the score that the "average American" (I.Q. 100) would be expected to make upon graduating from an average college.

2. He can compare his score on any test with the median scores of candidates taking any particular option by means of the profiles shown on the chart.

3. He can study his own *relative* achievements in different fields by direct visual inspection of his profile, since all tests are reported in Scaled Score units. This is perhaps the most important use of the chart, since it gives the candidate direct aid in appraising his own particular strengths and weaknesses.

It should be noted that many of the superintendents in the larger examining centers are likely to be interested in receiving applications and transcripts from candidates whose scores tend to be "exceptional," i.e., around 80 or above on the various examinations.

COOPERATIVE TEST SERVICE . . . 15 AMSTERDAM AVENUE, NEW YORK, N. Y.

TABLE II

INTERCORRELATONS OF SCORES ON THE TESTS OF THE 1940 TEACHER EXAMINATION BATTERY OBTAINED FROM A SAMPLE OF ONE-FIFTH OF THE CASES SELECTED TO BE REPRESENTATIVE OF ALL CENTERS AND ALL SCORE LEVELS (N = 668)

Test		1	2	3	4	5	6	7	8	9	10	11
Reasoning	1		.45	.45	.43	.34	.28	.46	.32	.67	.49	.28
English Comp.	2			.73	.75	.68	.76	.51	.73	.44	.72	.61
English Exp.	3				.51	.45	.67	.29	.59	.40	.57	.36
G.C.—Cur. Soc. P.	4					.81	.59	.58	.60	.47	.78	.78
G.C.—Hist. & Soc. St.	5						.64	.49	.58	.40	.70	.77
G.C.—Literature	6							.26	.76	.27	.57	.56
G.C.—Science	7								.37	.55	.59	.49
G.C.—Fine Arts	8									.23	.56	.37
G.C.—Mathematics	9										.55	.37
Prof. Information	10											.62
Cont. Affairs	11											
	M	60.43	64.37	61.76	62.44	62.98	63.34	60.59	63.52	62.05	62.87	63.32
	σ	11.78	11.30	12.56	11.74	11.95	11.77	11.48	11.55	11.68	12.02	11.06

Literature, Current Social Problems and Contemporary Affairs, and English Comprehension and English Expression turn out to do so, and other pairs show anticipated lack of agreement. The Reasoning Test, which was the only non-verbal test in the battery, shows very few substantial correlations. It is most closely associated with the Mathematics section of the General Culture Test. It is interesting to observe that the next largest correlation coefficient for this test is with the Professional Information Test. This is gratifying since an effort was made to introduce a substantial factor of professional judgment into the questions in the latter test.

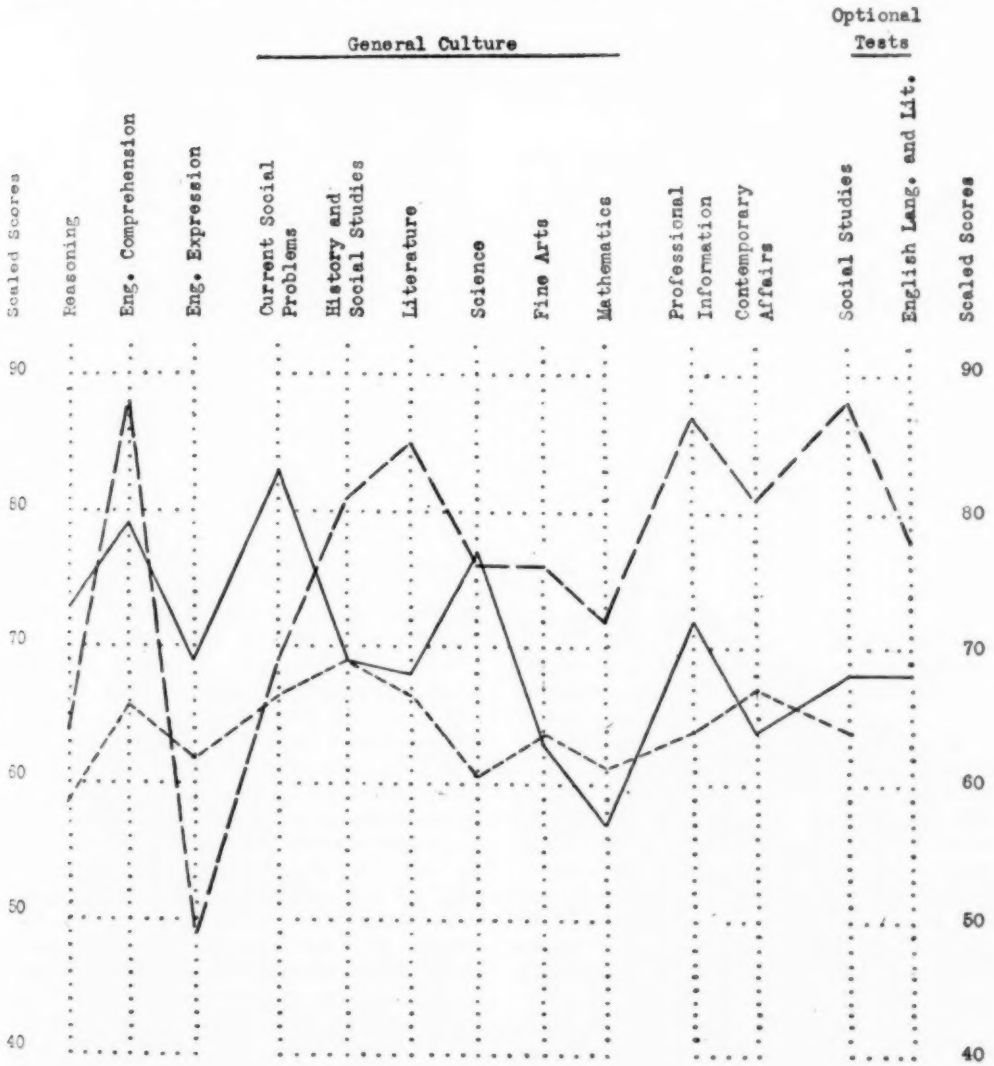
It is of some interest to note that the evidence from these intercorrelations and from the profiles for the subject-matter groups seems to suggest three main factors included in the various tests. Practically all the tests and especially those in the fields of English, Foreign Languages, Social Studies, Fine Arts, and Professional Information show high correlations with English Comprehension. This general agreement of tests in these fields has been observed in many previous studies and is usually considered to be due to a general verbal or linguistic factor which is important in practically all types of academic work. A second general factor which is indicated by the results appears to include mathematics, science, and non-verbal reasoning. Achievement in these fields seems to require something more than verbal ability and this something more is usually called a quantitative or reasoning factor. A third factor which is not quite so clear-cut as the two previously mentioned is related to the Current Social Problems, Contemporary Affairs, and Social Studies tests. This factor, consisting principally as it does of an awareness of social problems and current happenings, might be what some people have called "social sensitivity."

Although from inspection it appears that these three factors should account for most of the agreement in the scores on these tests, it should also be noted that there appears to be a moderately large factor which is unique or specific to each of the tests. These specific factors would be lost to a very large extent if the examination were greatly shortened or simplified. A more thorough mathematical analysis of the test results is planned and will be published in a later report.

One other type of evidence which favors separate testing and reporting for the various fields is obtained from a study of individual cases. In Figure 3 are shown the profiles of the scores made by two applicants for a position teaching the Social Studies in one of the cooperating cities. One of these candidates has a very good record on the History and Social Studies section of the General Culture Test and on three of the other sections of that test but is only slightly above the average of all the Social Studies applicants on the Current Social Problems part. The other candidate has a lower score on History and Social Studies and certain of the other sections of the General Culture Test but is outstanding on the Current Social Problems section. Other differences in the profiles of these candidates are apparent. One is much more outstanding on the Professional Information Test than is the other. These cases have been selected to show that although in general there is substantial agreement between the scores achieved on various tests such as the Current Social Problems and the History and Social Studies tests, there are individual ex-

FIGURE 3

PROFILES OF TWO CANDIDATES FOR A POSITION TEACHING THE SOCIAL STUDIES IN AN EASTERN CITY



Legend:

————Candidate A.
— — — —Candidate B.
----------Average for all candidates for Social Studies positions.

ceptions some of which are quite striking in character.

Numerous examples of cases of this kind might be cited, but it should be evident from the intercorrelations and other data presented that the task of filling a particular position is greatly simplified if such a comprehensive objective profile is available in addition to ratings of training and experience, personality characteristics, and other factors considered important by the local group.

There is another point in favor of a rather comprehensive examination that should be mentioned. Even if it developed that the forty-minute Reading Comprehension Test ranked the candidates in almost the same order as the total battery for this initial group, it would be very dangerous to abandon the use of the other tests. If it were known in advance that the test contained nothing but Reading Comprehension, there would be grave danger that other factors in the training of the candidates would be slighted, so that in later groups the Reading Comprehension Test would no longer provide the accurate indication of broad preparation which it originally did. For these reasons the National Committee plans to keep the battery of examinations as comprehensive as possible and has taken every possible opportunity to urge superintendents and boards of education to supplement the examinations so as to include as many of the important characteristics of good teachers as possible in the selection plan.

In addition to demonstrating that there are very large individual differences in the performance of candidates on these tests and also that the patterns of individuals whose total scores are at nearly the same level are frequently quite different, it seems important to present evidence showing as clearly as possible what these tests are measuring. The experiences of two of the cooperating superintendents will provide a type of evidence which should accumulate quite rapidly as the tests become more widely used. The first case concerns the report to us by the superintendent of a Western school system that an individual who had taken the examinations this March in an Eastern center had written to him applying for a position. The superintendent wrote to the National Committee office regarding the candidate's letter of application saying, "His use of the English language in form, in spelling, and in grammar was such that we had to tell him that we were not interested in any candidate of his type." Inspection of the records in the central office revealed that this particular candidate ranked close to the bottom among all candidates taking the examinations. He was in the lowest seventh of the distribution in both English Comprehension and English Expression.

The superintendent in another city where the examinations were used this past spring talked with a number of candidates who had taken the examinations and were applying for positions in his system. He reported to the National Committee that he found his judgment of these candidates substantially in agreement with the results which they obtained from the National Teacher Examinations. The National Committee plans to make a systematic survey this fall to discover the relationships between various factors usually associated with successful teaching and the scores achieved on the examinations. It is hoped that this survey will bring out more clearly those elements of successful teaching not adequately covered by these examinations so that the examinations may be extended, and so that school officials may be assisted in appraising those factors for which objective examinations are not now available.

In the meantime it is believed that the best evidence concerning the validity of the examinations is contained in the items themselves. Therefore, a number of items are presented below to enable the reader to judge for himself as to whether the examinations are providing valid measures of factors concerning which superintendents desire information. These particular items have been selected for presentation here because they provide the most valid brief indication of the contents of the test which is now available. In the case of most of these items the chances are about even that the average candidate will know the answer. But for selected persons obtaining scores within the top ten per cent on a particular test, the chances are generally about ten to one that they will know the answer to any of these sample questions from the test. These are the items which agree best with the scores on the various examinations, and therefore provide the best illustrations of the meanings of these scores.

It is apparent that the average candidate applying to a superintendent for a position will not usually know the answers to these questions. On the other hand, if superintendents or boards of education wish to employ

individuals who can answer such questions as these, they may insure this by selecting those candidates who in addition to having the other desirable qualities for the position, obtained high scores on these examinations. In looking over these questions from the common examinations required of all candidates the reader should remember that, although the items presented below include a very limited sample of the test, they illustrate the type of question which is known to the high-scoring individuals but is as likely as not to be unknown to the average applicant for a teaching position. It should also be remembered that, as indicated in Table I, there are quite a number of applicants for teaching positions in the city school systems included this year who know the answers to *hardly any* of these questions. These individuals, *all of whom have fulfilled the necessary requirements for teaching certificates and many of whom have actually obtained teaching positions,* get a few items correct, partly by chance, but when their scores are corrected for guessing, they are very close to zero.

In general the directions for these items ask for the selection of that one of the five choices given which *best* completes the statement. In a few cases the candidate is asked to find the one which is wrong.

ENGLISH COMPREHENSION, PART II,

It must not be supposed that the officials in charge of education desire the young to become educated. On the contrary, their problem is to impart information without imparting intelligence. Education should have two objects: first, to give definite knowledge; second, and more important, to create those mental habits which will enable people to acquire knowledge and form sound judgments for themselves. The first of these we may call information, the second intelligence. The utility of intelligence is admitted only theoretically, not practically: it is not desired that ordinary people should think for themselves, because it is felt that people who think for themselves cause administrative difficulties.

2. According to the writer, a truly educated population is regarded by the officials in charge as

1 theoretically undesirable. (23%, 2%)[4]
2 the hope of democratic society. (4%, *1%*)
3 the main objective of the schools. (3%, *1%*)
4 unnecessary. (5%, *2%*)
5 hard to manage. (65%, 94%)
omitted (0%, 0%) *not reached* (0%, 0%)

[4] After each answer the percentage of all candidates selecting that particular choice is given; the second figure (in italics) is the percentage of those individuals with scores in the highest ten per cent who chose that answer. The correct answers are printed in bold-face type.

6. The writer evidently considers that the most important function of education is to

1 impart useful information. (27%, *8%*)
2 provide vocational training. (0%, *0%*)
3 develop intelligent thinking. (69%, 91%)
4 create a stable social group. (3%, *1%*)
5 develop well-adjusted personalities. (1%, *0%*)
omitted (0%, 0%) *not reached* (0%, 0%)

Slade was a matchless marksman with a navy revolver. One morning at Rocky Ridge when he was feeling comfortable he saw a man approaching who had offended him some days before. "Gentlemen," said Slade, drawing, "it is a good twenty-yard shot—I'll clip the third button on his coat!" Which he did. The bystanders admired it. And they all attended the funeral, too.

7. Slade may best be described as

1 humorous. (3%, 0%)
2 cold-blooded. (44%, 70%)
3 revengeful. (34%, *19%*)
4 playful. (8%, 7%)
5 daring. (11%, *3%*)
omitted (0%, *1%*) *not reached* (0%, 0%)

8. The passage achieves its effect chiefly by means of

1 suspense. (13%, *10%*)
2 exaggeration. (5%, 0%)
3 direct quotation. (20%, *3%*)
4 matter-of-fact treatment. (60%, 86%)
5 detailed description. (2%, 0%)
omitted (0%, *1%*) *not reached* (0%, 0%)

ENGLISH COMPREHENSION, PART III, VOCABULARY,

84. *preponderant*

1 determined. (5%, *2%*)
2 preparatory. (0%, 0%)
3 advanced. (0%, *1%*)
4 thought out. (21%, *5%*)
5 predominant. (30%, 90%)
omitted (16%, *2%*) *not reached* (28%, *0%*)

ENGLISH EXPRESSION, GRAMMATICAL USAGE,

37. It was truly a magnificent hotel;
1 (4%, 0%)
every one of the washstands in the various guest rooms were of marble,
2 (64%, 96%)
as were the fireplace mantels.
3 (10%, 0%) 4 (0%, 0%)
5 *none considered wrong* (18%, 4%)
omitted (4%, 0%) *not reached* (0%, 0%)

ENGLISH EXPRESSION, PUNCTUATION,

All the childrens art classes were represented in
7
the exhibit, and the class which had done the best
8
work received a prize.
9

7. 1 No punctuation necessary. (10%, *1%*)
2 children's. (56%, 84%)
3 childrens'. (27%, *15%*)
omitted (5%, 0%) *not reached* (2%, 0%)

8. 1 **No punctuation necessary. (73%, 95%)**
   2 class, (18%, 4%)
   omitted (7%, 1%) *not reached* (2%, 0%)
9. 1 **No punctuation necessary. (65%, 94%)**
   2 work, (24%, 4%)
   omitted (9%, 2%) *not reached* (2%, 0%)

ENGLISH EXPRESSION, CAPITALIZATION,

"The truth is," she continued, "we almost forgot to mail them." 14

14. 1 capital letter. (19%, 8%)
    2 small letter. (67%, 92%)
    omitted (4%, 0%) *not reached* (10%, 0%)

ENGLISH EXPRESSION, SPELLING,

12. 1 **potatoe. (58%, 94%)**
    2 mattress. (2%, 0%)
    3 evidently. (1%, 0%)
    4 impromptu. (4%, 1%)
    0 none wrong. (18%, 4%)
    omitted (3%, 1%) *not reached* (14%, 0%)

GENERAL CULTURE, PART I, CURRENT SOCIAL PROBLEMS.

24. Freedom of speech is guaranteed to the American people by the
    1 unwritten law. (6%, 1%)
    2 penal code. (0%, 0%)
    3 first amendment to the Constitution. (63%, 98%)
    4 Zenger decision. (2%, 0%)
    5 Declaration of Independence. (28%, 1%)
    omitted (1%, 0%) *not reached* (0%, 0%)
55. A governor of a state has great power in labor disputes through his control over
    1 the court of labor relations. (10%, 0%)
    2 the expenditure of relief funds for strikers. (2%, 2%)
    3 the issuance of injunctions. (10%, 7%)
    4 the National Guard. (56%, 91%)
    5 interstate commerce laws. (10%, 0%)
    omitted (1%, 0%) *not reached* (11%, 0%)
60. In the past fifty years, man in the Western world has gained most in
    1 health. (52%, 97%)
    2 native intelligence. (6%, 0%)
    3 emotional stability. (4%, 0%)
    4 security of employment. (6%, 1%)
    5 stability of government. (14%, 2%)
    omitted (0%, 0%) *not reached* (18%, 0%)

GENERAL CULTURE, PART II, HISTORY AND SOCIAL STUDIES,

46. An innovation of the last ten years has been the United States government's encouragement of
    1 collective bargaining. (52%, 93%)
    2 the regulation of interstate commerce. (14%, 1%)
    3 the policy of appointment to government positions on the basis of merit. (8%, 0%)
    4 cooperative action to reduce prices. (6%, 1%)
    5 collective farming. (6%, 2%)
    omitted (12%, 3%) *not reached* (2%, 0%)
58. Neutrality is most often endangered by disputes arising in connection with
    1 the transport of soldiers. (2%, 1%)
    2 radio propaganda. (1%, 0%)
    3 seaborne commerce. (57%, 92%)
    4 interference in the internal policy of another nation. (19%, 3%)
    5 the public utterances of persons in high position. (4%, 4%)
    omitted (6%, 0%) *not reached* (11%, 0%)

GENERAL CULTURE, PART III, LITERATURE,

20. One of the greatest of French essayists was
    1 Ronsard. (2%, 0%)
    2 Montaigne. (54%, 98%)
    3 Molière. (13%, 2%)
    4 De Maupassant. (16%, 0%)
    5 Villon. (1%, 0%)
    omitted (13%, 0%) *not reached* (1%, 0%)
70. The poet who "sounded his barbaric yawp over the roofs of the world," yet brought a new freedom and flexibility to poetry, was
    1 T. S. Eliot. (4%, 4%)
    2 Emerson. (6%, 0%)
    3 Edgar Lee Masters. (4%, 0%)
    4 Edwin Markham. (4%, 0%)
    5 Walt Whitman. (55%, 95%)
    omitted (21%, 1%) *not reached* (6%, 0%)

GENERAL CULTURE, PART IV, SCIENCE,

9. Which one of the following is most directly affected by the moon?
    1 Radio communication. (2%, 0%)
    2 Proper planting periods. (16%, 0%)
    3 Fertility of animals. (2%, 0%)
    4 Probability of rain. (12%, 0%)
    5 Sea level (62%, 100%)
    omitted (6%, 0%) *not reached* (0%, 0%)
14. The feature which especially distinguishes man's nervous system from that of other vertebrates is the greater development in man of the
    1 respiratory center. (2%, 0%)
    2 nerve sheathing. (4%, 1%)
    3 sensory organs. (10%, 1%)
    4 spinal ganglia. (12%, 4%)
    5 forebrain. (67%, 94%)
    omitted (5%, 0%) *not reached* (0%, 0%)
60. The physical discomfort produced by a warm, badly ventilated room full of people is due mainly to
    1 lack of sufficient oxygen. (24%, 13%)
    2 accumulated carbon dioxide. (24%, 15%)
    3 poisonous vapors from people's bodies (4%, 1%)
    4 accumulated water vapor. (28%, 68%)
    5 the very low humidity of the air. (12%, 3%)
    omitted (0%, 0%) *not reached* (8%, 0%)

GENERAL CULTURE, PART V, FINE ARTS,

42. The best architecture is generally conceded to be that in which beauty is the natural expression of
    1 national feeling. (12%, 1%)
    2 function. (48%, 91%)

3 the style of the period. (13%, 1%)
4 strength. (13%, 6%)
5 a fine sense of decoration. (3%, 1%)
omitted (11%, 0%) *not reached* (0%, 0%)

76. What modern painters are largely responsible for the revival of mural painting today?

1 English. (5%, 0%)
2 German. (2%, 2%)
3 Italian. (12%, 3%)
4 Mexican. (53%, 91%)
5 French. (8%, 1%)
omitted (16%, 3%) *not reached* (4%, 0%)

GENERAL CULTURE, PART VI, MATHEMATICS,

15. Which one of the following equals 0.25%?

1 .0025. (53%, 95%)
2 .025. (6%, 0%)
3 .25. (34%, 5%)
4 2.50. (0%, 0%)
5 25.00. (4%, 0%)
omitted (3%, 0%) *not reached* (0%, 0%)

21. The positive square root of ¼ is

1 ½. (54%, 100%)
2 2. (4%, 0%)
3 $1\sqrt{2}$. (6%, 0%)
4 ⅛. (4%, 0%)
5 ⅜. (19%, 0%)
omitted (12%, 0%) *not reached* (1%, 0%)

PROFESSIONAL INFORMATION, PART I, EDUCATION AND SOCIAL POLICY,

1. The part the school is permitted to play in the development of society is determined chiefly by the

1 attitude of the society which maintains it. (67%, 87%)
2 moral courage of the teachers. (1%, 0%)
3 ability of the teachers to construct a program for social change. (19%, 7%)
4 power of judgment developed in the children. (11%, 5%)
5 willingness of the school to follow political domination. (0%, 0%)
omitted (2%, 1%) *not reached* (0%, 0%)

43. Substantial modifications in the high school curriculum have been necessary since the World War to allow for

1 expansion of the college preparatory curriculum. (14%, 5%)
2 the presence in high school of many persons with slight aptitude for book learning. (61%, 88%)
3 studies on a more difficult level, necessitated by better preparation in the grades. (3%, 3%)
4 emphasis on cultural rather than vocational subjects because of decreased employment opportunities. (5%, 2%)
5 consolidation of rural high schools. (10%, 2%)
omitted (7%, 0%) *not reached* (0%, 0%)

PROFESSIONAL INFORMATION, PART II, CHILD DEVELOPMENT AND EDUCATIONAL PSYCHOLOGY,

19. If a child is *continually* a difficult behavior problem, the teacher might best

1 refer his case to a competent authority for investigation. (71%, 90%)
2 discuss the problem with other teachers. (9%, 5%)
3 isolate him from the rest of the class. (2%, 0%)
4 have him transferred to another class. (2%, 0%)
5 devise a more effective form of discipline. (16%, 5%)
omitted (1%, 0%) *not reached* (0%, 0%)

55. Six-year-old children differ most from ten-year-olds in which one of the following respects?

1 Fine-muscle control. (48%, 80%)
2 Large-muscle control. (8%, 3%)
3 Visual acuity. (3%, 2%)
4 Auditory acuity. (1%, 0%)
5 Metabolic rate. (20%, 13%)
omitted (14%, 2%) *not reached* (6%, 0%)

PROFESSIONAL INFORMATION, PART III, GUIDANCE AND INDIVIDUAL AND GROUP ANALYSIS,

18. A boy applying for entrance to a highly selective school attained a percentile score equivalent to the 67th percentile of the previous year's entering class in that school. The school counselor is justified in concluding that the boy

1 probably should not be admitted to the school. (8%, 8%)
2 if admitted, would probably fail in his work. (2%, 0%)
3 if admitted, would have to work harder than the average entering student in order to avoid failure. (30%, 17%)
4 if admitted, would probably do slightly better than average work. (36%, 70%)
5 if admitted, should be about 67th best in the entering class of one thousand. (10%, 2%)
omitted (14%, 3%) *not reached* (0%, 0%)

26. It is a basic assumption of experts in vocational guidance that

1 interest in an occupation is the best evidence of future success in that occupation. (10%, 0%)
2 vocational aptitudes cannot be measured objectively. (4%, 2%)
3 pupils need assistance in estimating their own abilities, aptitudes, and interests. (84%, 98%)
4 there is one, and only one, occupation for which each individual is ideally suited. (0%, 0%)
5 vocational and educational guidance should be kept entirely separate. (1%, 0%)
omitted (1%, 0%) *not reached* (0%, 0%)

PROFESSIONAL INFORMATION, PART IV, SECONDARY SCHOOL METHODS,

1. The primary function of the recitation should be to
   1 give individual attention to the slower pupils. (1%, 0%)
   2 discover the pupils who have not assimilated previous lessons. (6%, 2%)
   3 discover the extent to which the individual pupils are able to recall material read. (1%, 0%)
   4 develop understanding through group discussion. (74%, 90%)
   5 improve the ability of the pupils to present material orally. (17%, 7%)
   omitted (1%, 1%) *not reached* (0%, 0%)

38. To be successful, the "no-failure" policy adopted in some schools requires
   1 the individualization of programs of study. (60%, 80%)
   2 decreasing attention to the superior students. (4%, 3%)
   3 the abandonment of school grades. (20%, 12%)
   4 emphasis on vocational education. (4%, 2%)
   5 elimination of pupils who cannot do satisfactory work. (4%, 2%)
   omitted (8%, 1%) *not reached* (0%, 0%)

SUMMARY AND CONCLUSIONS

The above discussion gives a brief analysis of the results of the first edition of the National Teacher Examinations. Those examinations were prepared under the direction of the staff of the Cooperative Test Service of the American Council on Education with the very generous cooperation and assistance of more than two hundred experts from all parts of the country. The examinations were designed to include those aspects of the candidates' training and ability about which the superintendents desired information and for which satisfactory techniques of objective measurement were available.

Only those items were included in the examination which performed satisfactorily in a preliminary tryout and which were selected and agreed to by a large number of experts and critics including subject-matter scholars, teacher training experts, and selected teachers and supervisors in city school systems. An effort was made to make the examinations as functional and valid as possible by stressing the *application* of facts and principles and the candidates' ability to exercise good judgment in selecting the best of a number of possible solutions of practical school problems.

The analysis of the results indicates that *candidates for teaching positions are not equally well qualified;* and that *a fairly long and varied battery of examinations is necessary to describe the various candidates adequately.* And finally, *evidence is presented to illustrate the validity of the various parts of the examinations in measuring their respective aspects of teacher preparation.*

In conclusion a word should be said about the use of these examination results. The National Committee established examining centers in about twenty of the larger cities for the initial administration of the tests in March, 1940. The National Committee has been interested to observe that in each of the more than twenty school systems requiring candidates to present transcripts of their records on the National Teacher Examinations, the selection procedure has also included an appraisal of training, experience, recommendations, and the more intangible personality characteristics. A further point of interest is that no two of the cities have adopted the same procedure. Some give the examinations a weight of about one-third, others one-half; some use the examinations as a preliminary qualifying hurdle, others require them only of persons who have been tentatively selected on the basis of preliminary interviews and ratings. But all of them require an interview and all seem to be making every effort to gain as much information as possible from persons with whom the candidate has studied or worked. A number of superintendents have stated that the availability of these objective and comparable examination records relieves them of the burden of interpreting ambiguous records of scholastic achievement and enables them to focus their attention more clearly on the appraisal of personality characteristics.

The National Committee on Teacher Examinations and the staff of the Cooperative Test Service wish to take this opportunity to acknowledge the very kind and generous cooperation of the many individuals in colleges and in public school systems who assisted in the preparation and administration of these examinations. Suggestions concerning the examinations and services provided by the National Committee are welcomed. It is hoped that, through the active support of the many persons interested in this project, it will be possible to make further improvements in the examinations and establish a larger number of examining centers for the convenience of candidates in the coming year.

SOME FACTORS IN PHYSICS ACHIEVEMENT AT THE COLLEGE LEVEL

D. B. Stuit and C. J. Lapp
University of Iowa

The prediction of scholastic success in various school subjects at the secondary and college level is one of our major personnel problems. Vocational guidance during this period consists quite largely of counseling with regard to the subjects which the student must take in order to prepare for entrance to a certain occupation. The student's previous scholastic record suggests itself as one of the best predictive indices of success in various curricula or subjects, but there are many occasions when the counselor has need for special measuring devices to supplement the student's record of scholastic attainments. Questions of differential prediction cannot be answered satisfactorily by an examination of school records alone.

A related personnel problem is that of adjusting instructional techniques to the student's particular level of ability. Granted that the student must take a certain subject, it is still possible to make adjustments in subject matter and techniques of teaching which will enable him to achieve satisfactorily. A student who under ordinary circumstances would be doomed to academic failure might conceivably do a satisfactory grade of work if his level of ability were to be taken into consideration. Likewise the student with high aptitude might gain a great deal more from a course if the subject matter proved sufficiently challenging. Hence, for both educational and vocational guidance purposes, aptitude tests for specific subjects or groups of subjects should prove valuable in clinical counseling.

The purpose of the present study was to identify certain fundamental factors related to scholastic success in college physics. For a number of years the Physics Aptitude Test of the Iowa Placement Examinations has been used at the University of Iowa to predict the scholastic success of students enrolled in physics courses. It has proved itself to be a satisfactory instrument for this purpose, particularly in the sectioning of students according to ability. Among the other tests measuring factors which seemingly were related to physics achievement, it occurred to the present writers that tests measuring the spatial relations factor might provide valuable data which would supplement the information provided by the Physics Aptitude Test. Three tests already available appeared to measure some of these abilities thought to be related to physics achievement, namely, the Minnesota Paper Form Board, the Mann Engineering Drawing Aptitude Test,[1] and the Mann Mutilated Cubes Test. Inasmuch as the majority of these students had taken the Iowa Qualifying Examination,[2] which includes the Iowa Mathematics Aptitude Test, the latter was included in the experimental test battery.

Since it was administratively difficult to require the entire experimental group of 155 students to take all the tests, it was broken up into two sections, Group I consisting of 80 students and Group II of 75. Both groups were given the Iowa Physics Aptitude Test and the Minnesota Paper Form Board. In addition to this battery, Group I was given the Mann Engineering Drawing Aptitude Test and Group II took the Mann Mutilated Cubes Test as the third component of the battery. The Iowa Mathematics Aptitude Test was available only for those students who had entered the University as freshmen and therefore a third group of 107 students was set up on the basis of these available tests.

During the first semester, the Cooperative Physics Achievement Tests (covering Mechanics, Heat, and Sound) were administered to all the students, and the total score, consisting of the sum of the scores on the subtests, was used as the criterion of achievement in physics. Prior to analyzing the predictive values of the various batteries, the

[1] Only parts 1, 3, and 7 of this test were used. The remaining parts do not deal with spatial relationships.

[2] The score in this examination consists of the sum of the weighted raw scores in the Iowa High School Content Examination, Iowa English Training Test, Iowa Mathematics Aptitude Test, and Iowa Silent Reading Test. The average correlation of this examination with first semester grade point averages is .65.

two groups were compared on the basis of general ability, as measured by the Iowa Qualifying Examination Composite Score, on physics aptitude as measured by the Physics Aptitude Test, and on physics achievement as measured by the Cooperative Physics Tests. No statistically significant differences were obtained.

The zero order inter-correlations, computed by the product-moment method, between the various aptitude tests and the criterion score for Group I are presented in Table I.

TABLE I

ZERO ORDER CORRELATIONS BETWEEN APTITUDE MEASURES AND CRITERION FOR GROUP I

N = 80

$r_{12} = .510$	$r_{23} = .662$
$r_{13} = .505$	$r_{24} = .535$
$r_{14} = .258$	$r_{34} = .160$

1. Cooperative Tests.
2. Mann Engineering Drawing.
3. Physics Aptitude.
4. Minnesota Paper Form Board.

Both the Mann Engineering Drawing Aptitude Test, $r_{12} = .510$, and the Physics Aptitude Test, $r_{13} = .505$, tend to correlate fairly high with the criterion score, whereas the magnitude of the correlation for the Minnesota Paper Form Board, $r_{14} = .258$, is considerably lower. The character of the other intercorrelations is worthy of note, and tends to suggest the existence of several factors underlying ability in physics. Whereas both the Physics Aptitude Test, $r_{23} = .662$, and the Minnesota Paper Form Board Test, $r_{24} = .535$, show a considerable degree of co-variability with the Mann Engineering Drawing Aptitude Test, the correlation of the Physics Aptitude Test with the Minnesota Paper Form Board tends to be quite small, $r_{34} = .160$. This may indicate that although the Minnesota Paper Form Board and the Mann Engineering Drawing both presumably measure ability to deal with spatial relations, the Mann Engineering Drawing test tends to be a better measure of those components related to physics aptitude and achievement.

Multiple correlations[3] were computed for the different combinations of aptitude tests correlated with the criterion score and these are presented below in descending order of magnitude of R.

$R_{1.234} = .588$	1. Criterion Score.
$R_{1.23} = .557$	2. Mann Engineering Drawing.
$R_{1.34} = .536$	3. Iowa Physics Aptitude.
$R_{1.24} = .510$	4. Minnesota Paper Form Board.

For Group I, the best multiple correlation was obtained for the combination of Mann Engineering Drawing Aptitude, Iowa Physics Aptitude and Minnesota Paper Form Board. This result seems in agreement with the previous inferences deduced from Table I concerning the apparently different components measured by the Mann Engineering Drawing and Physics Aptitude Tests. However, the engineering drawing and form board tests do not add greatly to the predictive power of the physics aptitude test used alone.

As previously noted, Group II was given the Mann Mutilated Cubes Test instead of the Mann Engineering Drawing, and in addition, the Physics Aptitude Test and the Minnesota Paper Form Board. The same type of correlational analysis used for Group I was repeated for Group II, the results being summarized in Table II.

TABLE II

ZERO ORDER CORRELATIONS BETWEEN APTITUDE MEASURES AND CRITERION FOR GROUP II

N = 75

$r_{12} = .349$	$r_{23} = .414$
$r_{13} = .664$	$r_{24} = .273$
$r_{14} = .261$	$r_{34} = .439$

1. Cooperative Tests.
2. Mann Mutilated Cubes.
3. Physics Aptitude.
4. Minnesota Paper Form Board.

It may be noted that even under slightly different conditions of testing, the correlation between the Minnesota Paper Form Board for Group II, $r_{14} = .261$, is of comparable magnitude to the correlation between these two components for Group I, $r_{14} = .258$. The Mann Mutilated Cubes Test does not seem to show as high a degree of co-variability with either the criterion score or the Physics Aptitude Test as did the Mann Engineering Drawing Aptitude Test. The correlation between the Physics Aptitude test scores and Minnesota Paper Form Board is considerably higher for Group II than for Group I. However, by the use of Fisher's Z-function the difference was found not to be significant at the 5% level. The same is true of the cor-

[3] These were computed by use of Griffin's technique. See Griffin, Harold D.: "Simplified schemas for multiple linear correlation," *Journal of Experimental Education*, I (March, 1933), pp. 239–254.

relation between the Physics Aptitude test scores and performance in the Cooperative Physics Achievement test.

The multiple correlations for the different combinations were determined and are presented below in descending order of magnitude.

$R_{1.23} = .673$ 1. Criterion Score.
$R_{1.234} = .670$ 2. Mann Mutilated Cubes Test.
$R_{1.34} = .665$ 3. Iowa Physics Aptitude.
$R_{1.24} = .376$ 4. Minnesota Paper Form Board.

The range of multiple correlations for Group II is much wider than in Group I, suggesting that the combination of the Mann Mutilated Cubes Test and Minnesota Paper Form Board is not as good a measure of the spatial factors related to physics achievement as the Mann Engineering Drawing Aptitude and Minnesota Paper Form Board. However, the physics aptitude test functioned more effectively as a predictive index for Group II than for Group I, and as a result all the multiple correlations involving it are higher. A small negative multiple regression coefficient resulted in a lower multiple correlation coefficient for four variables than for three.

The third group was made up of students from Groups I and II who had taken all the Cooperative Physics Achievement Tests, and for whom there were available test data on the Iowa Mathematics Aptitude and the Physics Aptitude tests. The zero order correlations are presented in Table III.

TABLE III

ZERO ORDER CORRELATIONS BETWEEN APTITUDE MEASURES AND CRITERION FOR GROUP III

N = 107

$r_{13} = .661$ 1. Cooperative Test Score.
$r_{23} = .657$ 2. Physics Aptitude Score.
$r_{12} = .779$ 3. Mathematics Aptitude Score.

The high correlation between Physics Aptitude and Mathematics Aptitude is very suggestive, and tends to indicate that the two tests measure either the same or highly correlated functions. The fact that one of the four subtests of the Physics Aptitude Test consists of problems in mathematics would, on subjective grounds, lead one to expect this result. The multiple correlation, $R_{1.23} = .699$, which is only slightly higher than either r_{12} or r_{13}, tends to lend further support to the hypothesis that these two tests are measuring very nearly the same thing.

Achievement test scores for semesters I and II were correlated for the two groups in order to determine the consistency of performance from semester to semester. For Group I the correlation coefficient is .774, for Group II it is .913, and for Group III it is .782, indicating a high degree of consistency of performance throughout the year.

During the year following the completion of the research described in the preceding paragraphs a supplementary study was made of the Minnesota Paper Form Board and the Mechanical Movements test of Thurstone's Primary Mental Abilities Test battery. One hundred ninety students participated in the study. The correlation between the form board scores and Cooperative Physics Achievement Test was found to be .152, thus confirming the findings of the previous study. The correlation for the Mechanical Movements test was .324. This was not as high as the correlations for the sub-tests of the Physics Aptitude Test as determined in another study, and consequently the test was not regarded by the present writers as measuring an important component of physics aptitude.

CONCLUSIONS AND INTERPRETATIONS

For the population included in this study or for similar populations, the following conclusions appear tenable:

1. Ability in mathematics appears to be more closely related than any other factor to achievement in college physics. While this cannot be regarded as an elemental component of physics aptitude, it appears to be operationally important.

2. The spatial relations factor as measured by the Minnesota Paper Form Board does not correlate highly with success in college physics.

3. Understanding of mechanical movements as measured by the Thurstone Primary Mental Abilities Tests does not bear a close relationship to success in college physics as measured in this investigation.

4. The Iowa Physics Aptitude Test and the Iowa Mathematics Aptitude Test predict success in college physics with a fairly high degree of accuracy.

FUNDAMENTAL PRINCIPLES IN TABULATING MACHINE METHODS OF STATISTICAL ANALYSIS

E. T. JOLLIFFE
Manager, Statistical Service, The State University of Iowa

FOREWORD

One of the major problems confronting the research worker is that of analyzing collected data in preparation for final interpretation and presentation. The method to be used for the analysis of the data for any type of research project is an individual question to be determined by the type of problem, the extent and form of the material at hand, the type of analysis to be made and the facilities and time available for making the analysis.

Because of the advantages in speed, accuracy and flexibility of tabulating machine analysis in statistical research, it is important that the research worker understand some of the fundamental principles involved in order that he may determine whether or not this method should be used, and if so, what preliminary steps should be taken to coordinate the collection of data with the method of analysis. Tabulating cards are physically movable records capable of being mechanically arranged, rearranged, counted, or tabulated, or a combination of these. Undoubtedly, the most satisfactory means of determining the method to be used is to consult an expert in the field of research project analysis, one who can evaluate the various methods of analysis in terms of the factors governing the method. If such an expert is not available, the research worker must make his own decision and do his own planning, and it is for the purpose of helping him with these problems that the following article is presented.

THE TABULATING CARD

The tabulating card in most common use measures 7⅜ inches by 3¼ inches and contains eighty columns of twelve positions each (Illustration 1). Ten of these positions are the printed digits 0 to 9 which correspond to the digits of numerical data to be punched. Two unprinted positions are at the top of the card and are known as the eleventh and twelfth positions, commonly referred to as X and R. Data are recorded in this card by means of punched holes.

For any project the card must be divided into fields to meet the requirements of that project. A field is a section of the card containing an assigned number of columns wherein a particular type of information is always to appear. For example, in Figure 2 (Example 1), columns 41–43 are reserved as the field for pupil number, columns 44–45 are reserved as the field for score. Likewise grade, age, and sex are assigned to their respective fields as shown. Any field must contain enough columns to accommodate the maximum number of digits that will be recorded in that field. Fields should be arranged so that punching can be accomplished with a minimum of effort and time. If the collection and recording of data has been properly planned, the arrangement of fields on the card will be identical with the arrangement of the data from which the cards are to be punched.

Any column of the card may contain more than one punched hole provided that only sorting and counting operations are to be performed on that column, or if it is desired to record alphabetical characters. Multiple punched columns will print from a tabulator only the lowest of the positions punched, except where provision is made for printing alphabetical characters. Additions can be performed only from the lowest positions punched. The X and R positions are useful only for sorting, counting and selection purposes, except when used in combination with another position to form an alphabetical character. They can be used neither for addition, nor indication on a tabulator.

CODING

Unless it is necessary to print a record of names or descriptions, all non-numerical information should be reduced to a numerical code. Each type of information, represented by the various fields on the tabulating card,

should be coded independently. In many projects data sheets may be constructed in such a manner that coding is largely or entirely automatic, insuring greater accuracy and saving of time for both the research worker and the punch operator.

Fractions should be avoided because there is no means of recording them on a tabulating card. All fractional parts should be recorded as decimal fractions.

Measurements to be algebraically manipulated should not be recorded in compound units, but reduced to a single unit. For example, feet and inches should be recorded as inches or as feet and tenths of feet. Age should be recorded in months rather than years and months.

If information from various sources must be transcribed to a final data sheet, the number of digits recorded for each variable should be the same as the number of columns reserved for that variable on the tabulating card. For example, a value of 96 for a variable that may have a maximum of three digits should be recorded as 096, or if four digits are a maximum, one being reserved for a decimal place, it should be recorded as 096.0. This rule, of course, can not usually be followed if the transcription to the tabulating card is made from the original data.

The Key Punch

By means of the electrical key punch, data are transcribed to the tabulating cards in the form of punched holes (Illustration 2).

The keyboard consists of twelve keys corresponding to the positions in a column of the tabulating card. As a key is depressed a hole is punched in the card, and the card advances to the next column much the same as a typewriter spaces as each character is printed (Illustration 3). Cards are automatically fed and ejected. Information that is common to a number of cards can be automatically punched into each card by placing a punched master card in the duplicating rack.

The Alphabetical Punch

The alphabetical punch contains all the features of the numerical key punch, and in addition has a typewriter keyboard for the purpose of punching alphabetical characters (Illustration 4).

When one of the letters on the typewriter keyboard is depressed, the combination of holes corresponding to that letter is automatically punched into the card. This machine is used to punch cards from which descriptions or names are to be printed (Illustration 5).

The Verifier

After the cards are punched, a second operator performs the same operations on the verifier that have been performed on the punch. Instead of punching holes, however, the verifier checks the accuracy of the holes already punched. Verification in such a manner insures accurate transcription from the data sheets to the tabulating cards (Illustration 6).

The Comparing Reproducer

The comparing reproducer is used to punch common information, numerical or alphabetical, into a group or groups of cards, eliminating the necessity of a punch operator repeating the same operation on each card of a group. It will also transfer punched data from one set of cards to another, either in the same field or in a different field. At the same time, or independently, the machine automatically checks its own punching. The operating speed is 100 cards per minute (Illustration 7).

The Sorting Machine

After the cards are punched, the sorting machine is used to arrange them in sequence or divide them into groups according to some classification or combination of classifications in order to make the needed tabulations. The sorting operation is based on the position of the punched hole in the column which is being sorted. On the machine there are twelve receiving pockets corresponding to the twelve digits on the card. As a card passes through the machine, it is directed to the pocket corresponding to the punched hole in the column being sorted by means of an electrical circuit closed by a brush contact through the punched hole. For example, a card punched 8 in the column being sorted would be directed to the "8" pocket; another card punched 4 in the same column would be directed to the "4" pocket (Illustration 8).

Sorting operations can be performed on only one column at a time; therefore, in orde

to arrange a group of cards in complete numerical sequence with respect to any field, it is necessary to sort each column of that field. A three column field would require running the cards through the sorter three times. Within any field, sorting should begin with the units column and progress through the tens, hundreds, thousands, etc. When sorting extends beyond a single field it should progress from the minor to the major fields. For example, if it is desired to arrange a group of cards in order by age within grade, it would be necessary to sort first on age and last on grade. For a group of cards punched like that shown in Figure 2 (Example 1), the sorting sequence would be column 49, column 48, column 47, and column 46. The machine operates at the rate of approximately 400 cards per minute.

The Card Counter

The card counter is an attachment to a sorter which performs the functions of counting (a single column at a time) the number of holes punched in each position in a column, and the number of cards. Sorting and counting can be accomplished simultaneously on any column. If sorting and counting are performed simultaneously for a group of cards on a column which contains no multiple punches, the counter will record the total number of cards, and the number of cards in each pocket. Even though the sorter will recognize only the lowest hole in a multiple punched column, the card counter will count each hole punched regardless of the number of holes in that column (Illustration 9).

The Tabulator

Printed reports are automatically prepared on the tabulator from the punched cards. As the cards pass through the machine, the punched holes actuate adding, subtracting, selecting, and printing mechanisms by means of electrical contacts. The arrangement of the compiled and printed reports is completely flexible. Therefore, the form of the printed reports is not dependent on the arrangement of fields on the tabulating cards.

The data on the printed reports can be automatically transferred to summary cards simultaneously with the printing of the reports. These summary cards can then be used for further analysis. The speed of the tabulator is approximately 150 cards per minute, modified to a certain extent by the number of lines and totals to be printed (Illustration 10).

In general there are two types of tabulators, numerical and alphabetical. Each will perform essentially the same operations except that the numerical tabulator will print only numerical data, whereas the alphabetical tabulator will print both numerical and alphabetical data.

Auxiliary Machines and Special Devices

Because of their relative unimportance in gaining a basic knowledge of the principles of tabulating machine methods of analysis, other auxiliary machines and special devices have not been presented. The principles of operation of such machines are in general much the same as those described, and they are adapted to perform automatically most of the routine statistical work in research problems.

Illustrative Problems

To give illustrative problems for all the different types of research projects that could be adapted to machine tabulation would be impossible. But, in order that a clearer understanding may be had of the manner in which some types of problems may be handled, a few examples of such problems are outlined in this section.

Each sample problem begins with an illustration of an original individual record sheet and a tabulating card punched from this record. A list of some of the possible tabulations that can be made follows, together with illustrative reports.

Example 1:

The objective test answer sheet shown in Fig. 1 has been marked by a student in answer to questions contained in a special test booklet. Fig. 2 shows a tabulating card punched from this answer sheet. The upper half of the first 39 columns of the card contains a punched hole record of the responses to the first 39 items, and the lower half, a record of the responses to the last 39 items. Responses 1, 2, 3, 4, and omit, respectively, are in positions R, X, 0, 1, and 3 in the upper half of the card, and in positions 4, 5, 6, 7, and 9 in the lower half. Hence, the responses to items 1 and 40 (3 and 2, respectively) are punched as 0 and 5 in column 1; responses

to items 2 and 41 (1, and 4, respectively) are punched as R and 7 in column 2. The R in column 40 is duplicated to all cards and serves as a check for the card punch operator in determining when the first 39 items have been completed. The card is returned to column 1 before the punching of the last 39 items is begun.

The following is a table showing the card fields in which each of the remaining types of data are punched.

Card Columns	Data Recorded
41–43	Identification number
44–45	Score
46–47	Grade
48–49	Age
50	Sex (1 = boy; 2 = girl)

Some of the possible tabulations are: 1—item analysis; 2—frequency distribution of scores; 3—sums of scores and numbers of pupils; 4—sums of squared scores; 5—correlation of odds and evens for reliability (scoring to be done with card stencils). Each of the above tabulations can be made in total or in groups by achievement, grade, sex, age, or responses to items or groups of items. Such tabulations contain the basic data needed for computing the desired statistical measures.

Fig. 3 is an analysis of six items of the test answer sheet shown in Fig. 1, by achievement groups for 300 pupils. The upper, middle, and lower thirds are determined by grouping in accordance with the total score on the test. In this case, approximately one third of the pupils scored above 33, one third between 27 and 33, and one third below 27 (see Fig. 6). In order to simplify the computations the groups were adjusted at each extremity so that there were 100 cases in each group. This analysis shows that a total of 128 pupils indicated response No. 1 as the answer to item No. 1, 73 of whom are in the upper third of the distribution of scores, 40 in the middle third, and 15 in the lower third. Also, 126 pupils indicated response No. 4 as the answer to the same item, 25 of whom are in the upper third, 45 in the middle third and 56 in the lower third. Likewise, a total of 105 pupils indicated response No. 2 as the answer to item 41, 29 of whom are in the upper third, 32 in the middle third, and 44 in the lower third. Item 1 and item 40 are on the same line because they are both punched in column 1, and are counted simultaneously.

Fig. 4 is an analysis of the same six items by grade. Of the 128 pupils indicating response No. 1 for item No. 1, 14 are in the eleventh grade, and 114 are in the twelfth grade. In terms of percentages, 23.0% of the eleventh graders, and 96.5% of the twelfth graders made this response.

Fig. 5 is a cross-analysis between item 1 and item 2. Of the 126 pupils selecting response No. 4 to item No. 1, 118 selected response No. 1 to item No. 2, one selected response No. 2, 3 selected response No. 3, 2 selected response No. 4, and 2 omitted the item. Likewise, of the 278 that selected response No. 1 to item No. 2, 118 selected response No. 1 to item No. 1, 12 selected response No. 2, 8 selected response No. 3, 12 selected response No. 4, and 22 omitted the item.

The first column in Fig. 6 is an indication of the test scores in ascending order. The other six columns are, as labelled, the frequency and cumulative frequency for the boys, girls, and total group.

Figs. 7 and 8 are sums of scores and numbers of cases for each response to item 1 and for age respectively. The averages have been computed with a calculating machine.

Fig. 9 contains the cumulative sums of scores by sex for each of the separate digits in the units and the tens positions of the score. The last number (5079 for boys) in the cumulative column is the sum of all the scores. The sum of the squared scores (ΣX^2) is obtained by adding the cumulative sums for each of the digits, except zero, in the units position and ten times the cumulative sums of each of the digits, except zero, in the tens position. This and the remaining figures are obtained on a calculating machine and represent the computations for the mean and stand deviation from the raw-score formulas: $M = \frac{\Sigma X}{N}$ $SD = \sqrt{\frac{\Sigma X^2 - M^2}{N}}$ By combining sums of scores and sums of squared scores for boys and girls, a mean and standard deviation can be computed for the total group.

Fig. 10 is similar to Fig. 9, except that each score has been divided into two parts—the score on odd items and the score on even items—and cumulative sums obtained for each. In addition, the cumulative sums have been obtained for the odd-item scores (X)

and each digit in the units and tens positions of the even-item scores (Y); so that now we have, by adding the column labelled "X" for the tabulation indicating digits for Y, the sum of the cross-products (ΣXY) between X and Y. Using a calculating machine, the standard deviations are computed as in Fig. 9, and the Pearson product-moment coefficient of correlation is computed from the raw-score formula:

$$r = \frac{\frac{\Sigma XY}{N} - M_x M_y}{SD_x \cdot SD_y}$$

Example 2:

Fig. 11 shows two pages from a questionnaire that is a good example of one which is self-coding. Parentheses indicate the codes and the card column numbers for each item of information. Fig. 12 illustrates the tabulating card punched from these two pages.

Any type of analysis or cross-analysis of the information contained in the questionnaire can be obtained from the cards. A few sample illustrations are shown in Figs. 13, 14, 15.

Fig. 13 is a distribution by age (column 1), and is interpreted as follows: 5 persons did not answer the item of age, one is from 18–19 years of age, 13 are from 20–21 years of age, 173 are from 20–24 years of age, etc.

Fig. 14 is a distribution of items 3, 4, and 5 (column 3), and is interpreted as follows: Of the 989 cases, 509 are men, and 480 are women; 473 are single, 469 are married, 16 are widows or widowers, and 12 are divorced. The early environment of 214 cases was the farm, of 157 cases the small town, of 238 cases the cities from 2000–15,000 population, and of 360 cases the cities over 15,000 population.

Fig. 15 is a cross-analysis of age (column 1) versus present population (column 7). Of the 173 cases from 22–24 years of age, one person is teaching in a city with a population of from 1000–1499, one is teaching in a city with a population of from 1500–1999, 4 are teaching in cities with populations of from 2000–2999, etc.

Example 3:

Fig. 16 is a sample of a data sheet to which has been posted a record of twelve measures for each individual case. Fig. 17 is a tabulating card punched from case No. 1. Note that numbers having fewer digits than have been allowed for in the tabulating card field are preceded by zero.

From these cards can be tabulated various distributions, sums, and progressive totals by digits for computing sums of squares and sums of cross-products for use in the formula for the coefficient of correlation. Such a tabulation of progressive totals together with the computations for the correlation coefficient are shown in Fig. 18 for a sample of 100 cases.

The column to the extreme right of each series of tabulations is used for the sums of squares to obtain the standard deviation of each variable; and all other columns are used for the cross-products between the column that is being added and the column to the extreme right. This tabulation follows the same principle as that in Fig. 9, except that columns for 11 cross-products are obtained simultaneously. Cross-products may be obtained between one variable and as many as 20 other variables in one set of cumulative totals. All computations have been performed on a calculating machine.

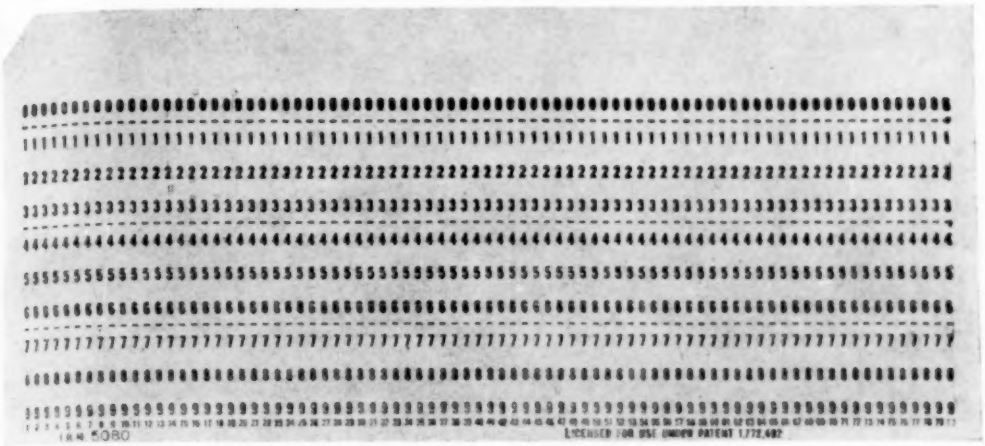


Illustration 1. The Tabulating Card.

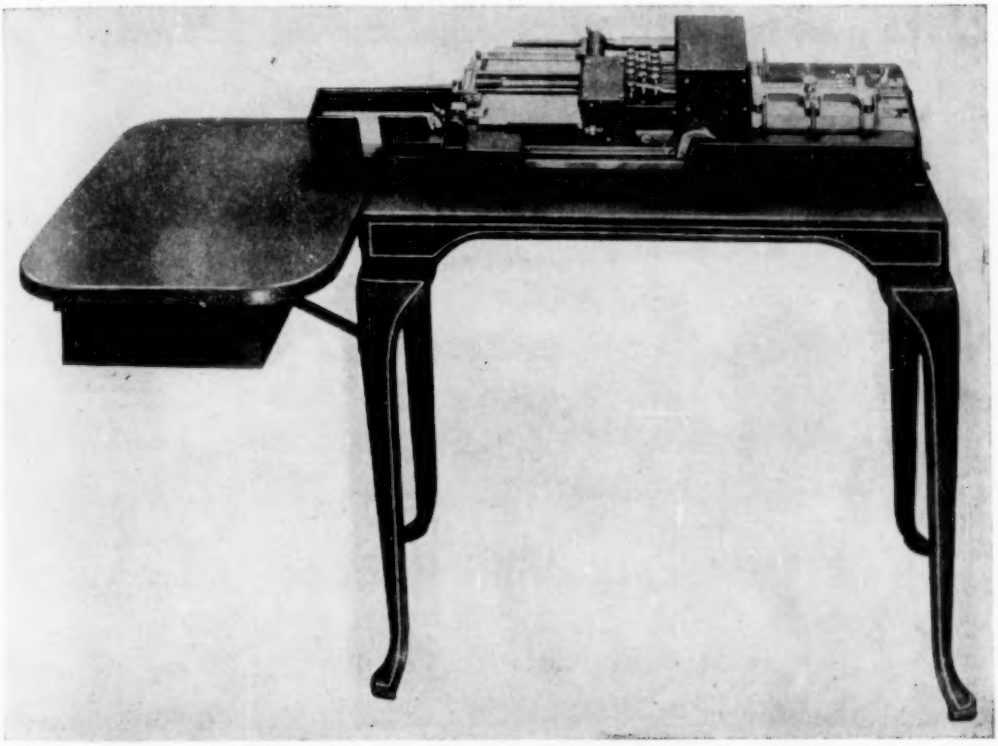


Illustration 2. International Duplicating Key Punch.

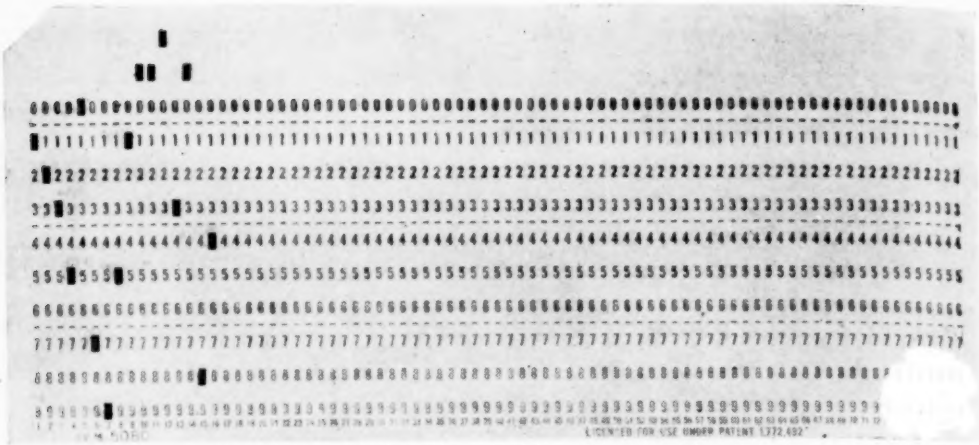


Illustration 3. Tabulating Card Punched on a Key Punch.

Illustration 4. International Alphabetic Duplicating Key Punch.

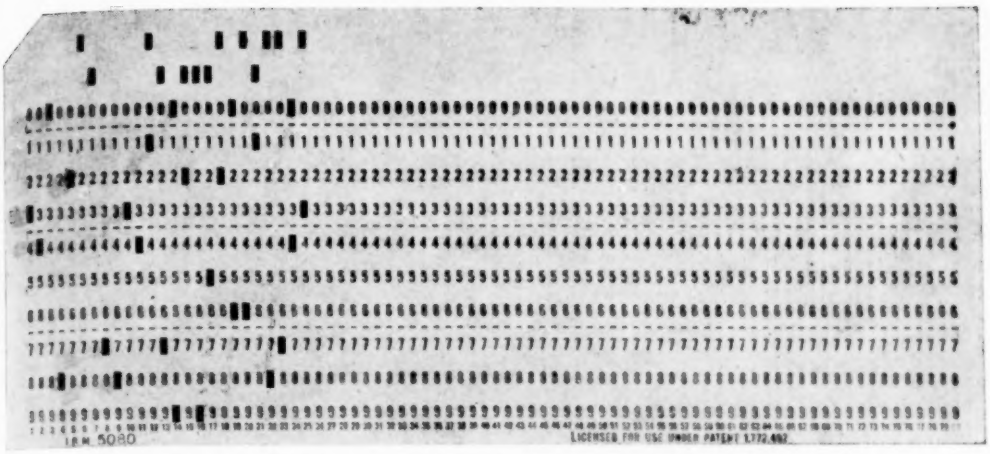


Illustration 5. Tabulating Card Punched on an Alphabetical Key Punch

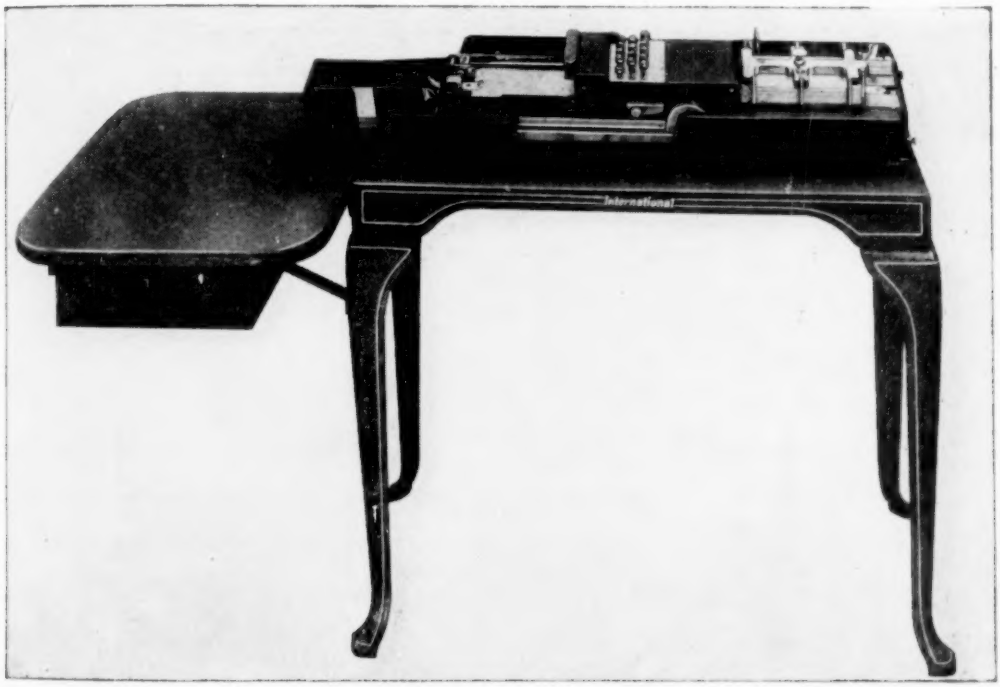


Illustration 6. International Motor Drive Verifier.

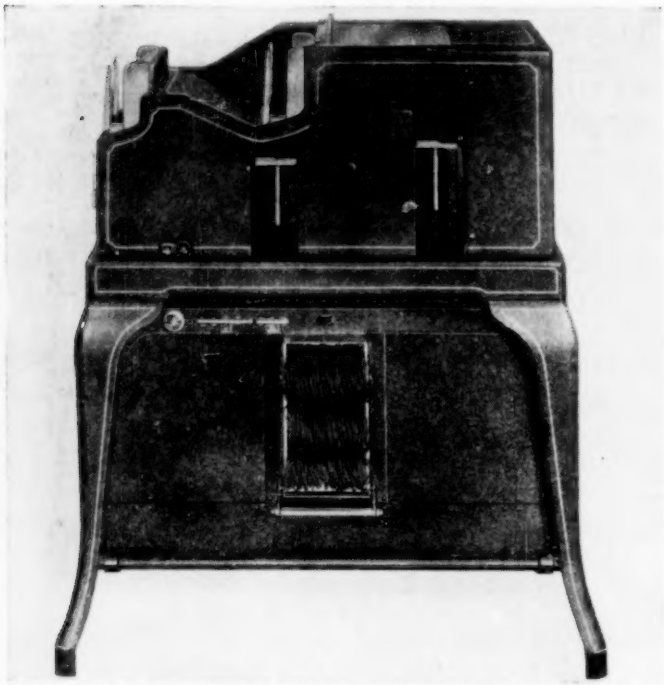


Illustration 7. International Comparing Reproducing Punch.

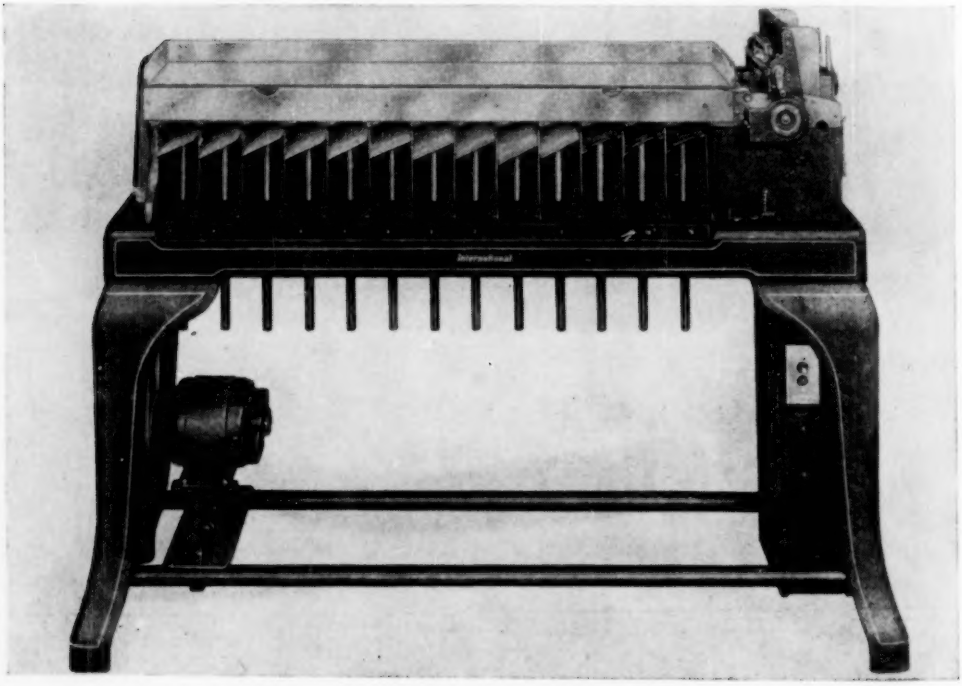


Illustration 8. International Horizontal Sorter.

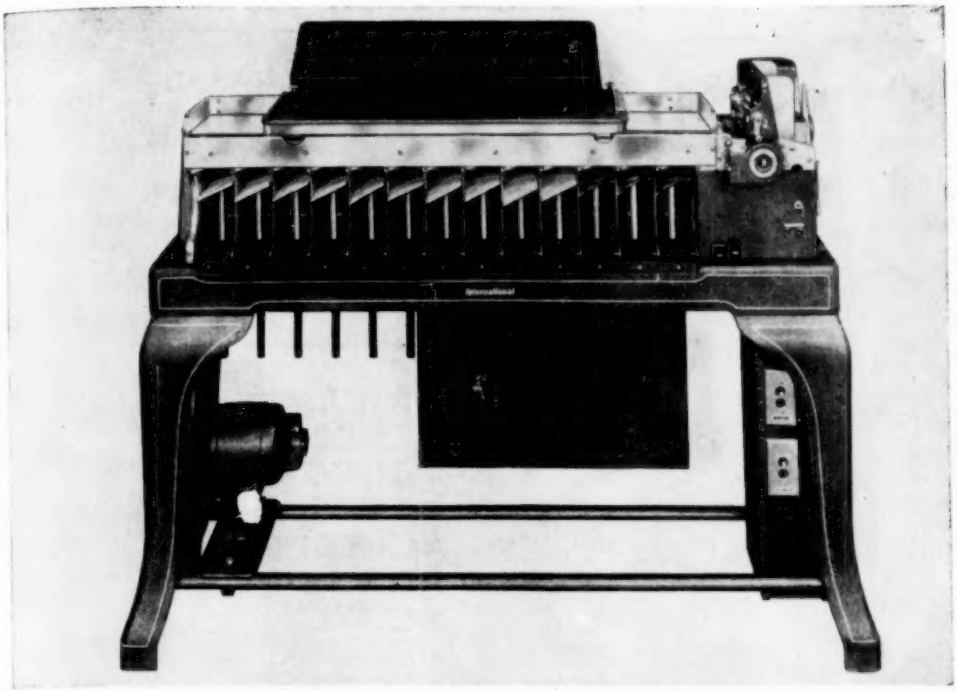


Illustration 9. International Card Counting Sorter.

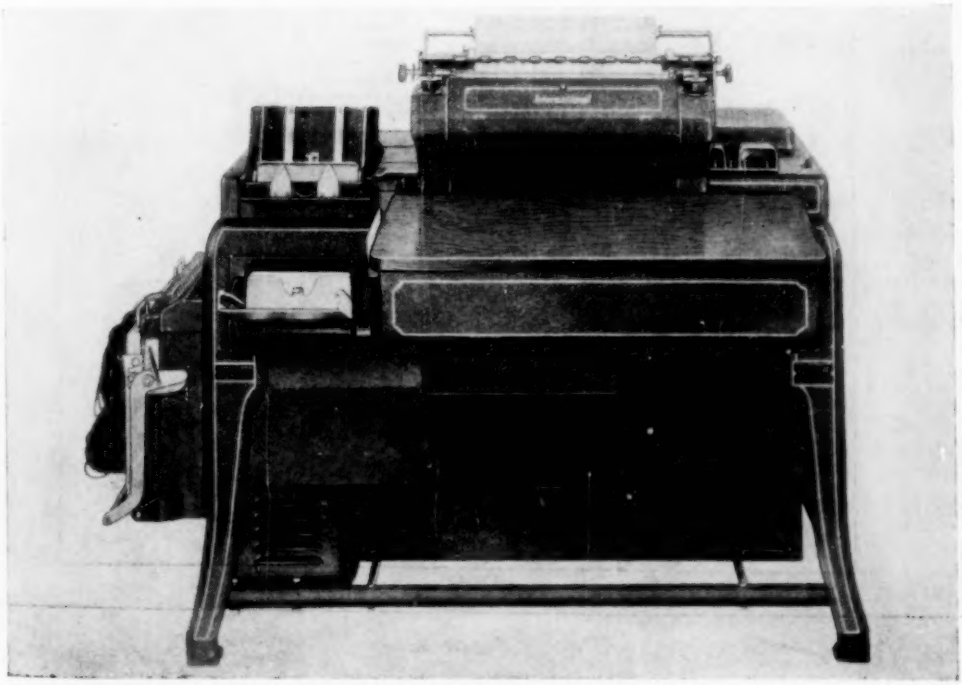


Illustration 10. International Alphabetic Accounting Machine Type 405.

274

ANSWER SHEET: AMERICAN GOVERNMENT

Name ..... Jones ..... John ..... E ..... Score 57
Last name first name initial

Grade: 11 Age: 15 yrs. Check: Boy ✓ Girl ....

Name of teacher: Miss Adams

School: Public High School

City: Anytown

Sample: 0 ☐☐☒☐	25 ☐☐☒☐	50 ☒☐☐☐	75 ☒☐☐☐
1 ☐☐☒☐	26 ☐☐☒☐	51 ☐☒☐☐	76 ☐☐☐☒
2 ☒☐☐☐	27 ☐☐☐☒	52 ☐☐☒☐	77 ☐☒☐☐
3 ☐☒☐☐	28 ☐☐☐☒	53 ☐☒☐☐	78 ☒☐☐☐
4 ☒☐☐☐	29 ☐☐☐☒	54 ☐☒☐☐	
5 ☒☐☐☐	30 ☐☒☐☐	55 ☐☐☐☒	
6 ☐☐☐☒	31 ☐☐☒☐	56 ☐☐☒☐	
7 ☐☒☐☐	32 ☒☐☐☐	57 ☐☐☐☒	
8 ☐☐☒☐	33 ☐☒☐☐	58 ☐☐☐☒	
9 ☐☐☒☐	34 ☒☐☐☐	59 ☐☐☐☒	
10 ☐☐☒☐	35 ☐☒☐☐	60 ☐☒☐☐	
11 ☒☐☐☐	36 ☐☐☒☐	61 ☒☐☐☐	
12 ☐☒☐☐	37 ☐☒☐☐	62 ☒☐☐☐	
13 ☐☒☐☐	38 ☐☐☐☒	63 ☐☐☒☐	
14 ☐☐☐☒	39 ☒☐☐☐	64 ☐☒☐☐	
15 ☒☐☐☐	40 ☐☒☐☐	65 ☐☒☐☐	
16 ☐☒☐☐	41 ☐☐☐☒	66 ☐☒☐☐	
17 ☐☐☐☒	42 ☐☐☒☐	67 ☐☒☐☐	
18 ☐☐☒☐	43 ☐☐☒☐	68 ☐☒☐☐	
19 ☐☒☐☐	44 ☐☒☐☐	69 ☐☐☒☐	
20 ☐☒☐☐	45 ☐☒☐☐	70 ☐☒☐☐	
21 ☐☒☐☐	46 ☐☒☐☐	71 ☒☐☐☐	
22 ☒☐☐☐	47 ☐☒☐☐	72 ☐☐☒☐	
23 ☒☐☐☐	48 ☒☐☐☐	73 ☐☐☐☒	
24 ☒☐☐☐	49 ☐☒☐☐	74 ☐☐☐☒	

Figure 1.

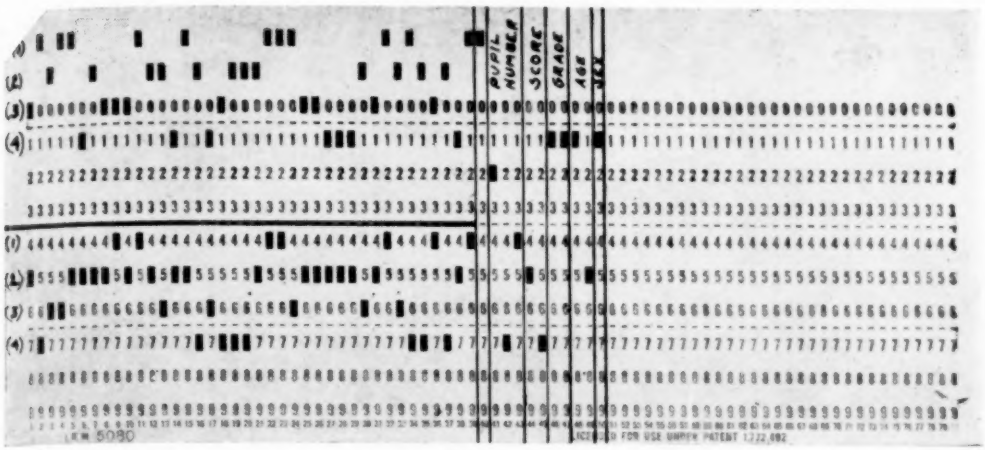


Figure 2.

Item	Group	Response 1	2	3	4	Omit	Item	Response 1	2	3	4	Omit	Total
1	Upper	73	------	1	25	1	40	71	2	14	9	4	100
	Middle	40	5	4	45	6		56	7	22	7	8	100
	Lower	15	8	4	56	17		42	7	24	13	14	100
	Total	128	13	9	126	24		169	16	60	29	26	300
2	Upper	95	3	1	1	------	41	3	29	10	55	3	100
	Middle	90	3	4	2	1		5	32	13	41	9	100
	Lower	93	------	5	------	2		8	44	14	27	7	100
	Total	278	6	10	3	3		16	105	37	123	19	300
3	Upper	58	15	13	3	11	42	54	2	14	25	5	100
	Middle	32	11	11	9	37		32	3	21	32	12	100
	Lower	17	14	6	13	50		18	7	21	36	18	100
	Total	107	40	30	25	98		104	12	56	93	35	300

Figure 3.

Item	Grade	Response 1	2	3	4	Omit	Item	Response 1	2	3	4	Omit	Total
1	11	14	4	------	31	12	40	28	7	13	6	7	61
	12	114	9	9	95	12		141	9	47	23	19	239
	Total	128	13	9	126	24		169	16	60	29	26	300
2	11	57	1	------	2	1	41	2	27	10	18	4	61
	12	221	5	10	1	2		14	78	27	105	15	239
	Total	278	6	10	3	3		16	105	37	123	19	300
3	11	22	7	5	2	25	42	17	2	8	20	14	61
	12	85	33	25	23	73		87	10	48	73	21	239
	Total	107	40	30	25	98		104	12	56	93	35	300

Figure 4.

ITEM 2

Response	1	2	3	4	Omit	Total
1	118	4	5	1	------	128
2	12	------	1	----------		13
3	8	------	1	----------		9
4	118	1	3	2	2	126
Omit	22	1	----------		1	24
Total	278	6	10	3	3	300

Figure 5.

	Boys		Girls		Total	
Score	F	C. F.	F	C. F.	F	C. F.
12	----------		1	1	1	1
15	1	1	------	1	1	2
17	------	1	2	3	2	4
19	3	4	3	6	6	10
20	3	7	3	9	6	16
21	1	8	4	13	5	21
22	4	12	9	22	13	34
23	3	15	10	32	13	47
24	5	20	5	37	10	57
25	7	27	9	46	16	73
26	8	35	12	58	20	93
27	7	42	5	63	12	105
28	9	51	10	73	19	124
29	3	54	7	80	10	134
30	6	60	9	89	15	149
31	8	68	8	97	16	165
32	12	80	8	105	20	185
33	7	87	1	106	8	193
34	6	93	7	113	13	206
35	8	101	4	117	12	218
36	3	104	3	120	6	224
37	2	106	5	125	7	231
38	10	116	4	129	14	245
39	4	120	1	130	5	250
40	3	123	2	132	5	255
41	4	127	4	136	8	263
42	------	127	3	139	3	266
43	1	128	1	140	2	268
44	3	131	1	141	4	272
45	6	137	2	143	8	280
46	2	139	------	143	2	282
47	1	140	------	143	1	283
48	1	141	------	143	1	284
50	1	142	1	144	2	286
51	1	143	2	146	3	289
52	1	144	1	147	2	291
54	1	145	------	147	1	292
55	1	146	------	147	1	293
56	1	147	1	148	2	295
59	1	148	------	148	1	296
61	1	149	------	148	1	297
65	------	149	1	149	1	298
66	1	150	------	149	1	299
73	1	151	------	149	1	300

Figure 6.

Responses on Item 1	Sum of Scores	No. of Cases	Average
1	4746	128	37.1
2	345	13	26.5
3	259	9	28.8
4	3610	126	28.7
Omit	610	24	25.4

Figure 7.

Age	Sum of Scores	No. of Cases	Average
xx	697	23	30.3
14	23	1	23.0
15	267	9	29.7
16	2021	62	32.6
17	4351	134	32.5
18	1749	56	31.2
19	387	13	29.8
20	75	2	37.5

Figure 8.

	TENS	UNITS	CUM SUM OF SCORES	Average
BOYS		9	359	
		8	1039	
		7	1349	
		6	1879	
		5	2674	
		4	3184	
		3	3600	
		2	4124	
		1	4669	
			5079	
	7		73	37.1
	6		200	26.5
	5		577	28.8
	4		1493	28.7
	3		3738	25.4
	2		5007	
	1		5079	
			5079 N = 151	
Σx^2			184547	
$\Sigma x^2/N$			1222.1656	
M.			33.6358	
S.D.			9.5288	

	TENS	UNITS	CUM SUM OF SCORES	Average
GIRLS		9	299	
		8	731	
		7	1085	
		6	1561	
		5	2081	
		4	2483	
		3	2789	
		2	3433	
		1	4031	
			4491	
	6		65	30.3
	5		325	23.0
	4		872	29.7
	3		2541	32.6
	2		4388	32.5
	1		4491	31.2
			4491 N = 149	29.8
Σx^2			145313	37.5
$\Sigma x^2/N$			975.2550	
M.			30.1409	
S.D.			8.1720	

Figure 9.

	TENS	UNITS	ODD SCORES (X)	EVEN SCORES (Y)
		9	351	390
		8	777	840
		7	1032	1126
		6	1472	1622
		5	1769	1927
		4	2147	2297
		3	2644	2851
		2	3146	3427
		1	3571	3875
N = 300			4181	4485
		3	196	220
		2	1194	1324
		1	3851	4187
			4181	4485
$\Sigma y'$ or Σxy			69319	75665
Σ/N			231.0633	252.2167
M				14.9500
SD				5.3576
$SD_x \cdot SD_y$			29.7477	
r			.7634	

	TENS	UNITS	
		9	373
		8	703
		7	1167
		6	1389
		5	1954
		4	2272
		3	2796
		2	3380
		1	3841
			4181
	3		126
	2		1060
	1		3778
			4181
Σx^2			67515
Σ/N			225.0500
M			13.9367
SD			5.5514

Figure 10.

Applies to all. Indicate with a cross (x) the age group below which best represents your age in years to your nearest birthday:

(R) ..... 18 to 19 years
(X) ..... 20 to 21 years
(0) ..... 22 to 24 years
(1) ..... 25 to 27 years
(2) ..X.. 28 to 30 years
(3) ..... 31 to 33 years
(4) ..... 34 to 36 years
(5) ..... 37 to 40 years
(6) ..... 41 to 45 years
(7) ..... 46 to 50 years
(8) ..... 51 to 60 years
(1)—(9) ..... 61 years and over

2. Applies to all. Indicate with a cross (x) the age group below which best represents your age at the time you entered the teaching profession:

(R) ..... 15 to 17 years
(X) ..... 18 to 19 years
(0) ..... 20 to 21 years
(1) ..X.. 22 to 24 years
(2) ..... 25 to 27 years
(3) ..... 28 to 30 years
(4) ..... 31 to 33 years
(5) ..... 34 to 36 years
(6) ..... 37 to 40 years
(7) ..... 41 to 45 years
(8) ..... 46 to 50 years
(2)—(9) ..... 51 years and over

3. Applies to all. Indicate with a cross (x) your sex:

(0) ..X.. male
(1) ..... female

4. Applies to all. Indicate with a cross (x) your social status:

(2) ..... single
(3) ..X.. married
(4) ..... widow or widower
(5) ..... divorced

5. Applies to all. Indicate with a cross (x) your early environment:

(6) ..X.. farm
(7) ..... small town (under 2000)
(8) ..... city (2000 to 15000 pop.)
(3)—(9) ..... city (over 15000 pop.)

6. Applies to all. Indicate with a cross (x) your religious preference:

(X) ..... Roman Catholic
(0) ..... Presbyterian
(1) ..X.. Methodist
(2) ..... Baptist
(3) ..... Lutheran
(4) ..... Congregational
(5) ..... Episcopalian
(6) ..... Other Protestant sects
(7) ..... Jewish

7. Applies to all. If you have no religious preference, put a cross (x) in the space below:

(8) ..... no preference

8. Applies to all. If you are a *member* of the sect indicated above, put a cross (x) in the blank below:

(4)—(9) ..X.. church member

2

Figure 11. (1st page)

9. Applies to all. In the list below, indicate with a cross (x) your father's occupation while living: Note: Check but one item.

(R) X farmer, agriculturist

(X) agent, office worker, salesman, realtor

(0) banker, financial agent

(1) contractor, building trades: carpenter, bricklayer, plumber, etc.

(2) government and public service

(3) merchant, dealer, proprietor

(4) professions: doctor, lawyer, dentist, teacher, preacher, etc.

(5) railroad employee

(6) mining (laborer)

(7) trades other than building, (skilled mechanic, painting, decorator, etc.)

(8) manufacturer, public utilities

(6) (9) unskilled laborer

10. Applies to all. In the list below, indicate with a cross (x) the type of teaching certificate you now have registered with your county superintendent: (If you hold more than one certificate, indicate only the highest.)

(R/0) X First Grade State

(R/1) Second Grade State

(R/2) Third Grade State

(R/3) Fourth Grade State

(R/4) Rural School Teacher's State

(R/5) Special State (Kindergarten)

(R/6) Special State (Primary)

(R/7) Other Special State

(X/0) First Grade County Uniform

(X/1) Second Grade County Uniform

(X/2) Third Grade County Uniform

(X/3) Special County

(X/4) Provisional County

(X/5) High School Normal Training

(X/6) Life Diploma

The above list should contain *only one* check mark.

11. Applies to all. If the certificate which you indicated in the preceding list was granted you by examination, place a cross (x) below:

(8) X certificate granted by examination

12. Applies only to holders of *life* certificates. If the certificate you have checked in the preceding list is validated for life, place a cross (x) below:

(6) (9) life certificate

13. Applies only to holders of *special* certificates. If you have indicated a *special* certificate, write below the name of the special subject you are certified to teach:

_____

14. Applies to all except rural teachers. If you are *not* teaching in a rural school, place a cross (x) before the population group to which the city in which you are *now teaching* belongs:

(R) below 500

(X) 500-999

(0) 1000-1499

(1) 1500-1999

(2) X 2000-2999

(3) 3000-3999

(4) 4000-4999

(5) 5000-9999

(6) 10000-14999

(7) 15000-24999

(8) 25000-49999

(7) (9) 50000 and over

15. Applies to all who have taught in more than one community. If you have taught in a town or city other than the one in which you are now teaching, indicate with a cross (x) the population group to which the city belongs in which you taught *just before* coming to your present position:

(R) below 500
(X) 500-999
(0) 1000-1499
(1) X 1500-1999
(2) 2000-2999
(3) 3000-3999
(4) 4000-4999
(5) 5000-9999
(6) 10000-14999
(7) 15000-24999
(8) 25000-49999
(8) (9) 50000 and over

2

Figure 11. (2nd page)

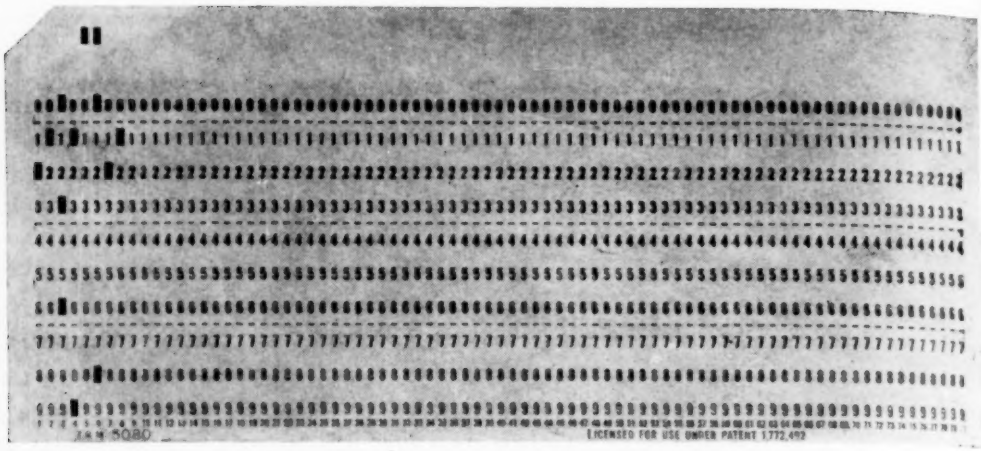


Figure 12.

Age (Col. 1)	Unans.	R	X	0	1	2	3	4	5	6	7	8	9	Total
	5	1	13	173	182	181	131	105	75	37	31	39	16	989

Figure 13.

	0	1	2	3	4	5	6	7	8	9	Total
Col. 3	509	480	473	469	16	12	214	157	238	360	989

Figure 14.

Present Enrollment (Col. 7)

Age (Col. 1)	Unans.	R	X	0	1	2	3	4	5	6	7	8	9	Total
Unans.								1	1	2		1		5
R												1		1
X	1						1		6	3	1		1	13
0	1			1	1	4	2	7	54	36	10	21	36	173
1	2						2	7	46	35	13	21	56	182
2	1			1		2	4	3	36	36	15	26	57	181
3	2	1				1	5	6	24	22	12	17	41	131
4		1			1		1	5	18	20	6	20	33	105
5	1					2	1	1	13	11	2	18	26	75
6								3	2	6	1	5	20	37
7									3	8	1	8	11	31
8						1		3	4	4	2	4	21	39
9							1		4	2	1	3	5	16
Total	**8**	**2**		**2**	**2**	**10**	**17**	**36**	**211**	**185**	**64**	**145**	**307**	**989**

Figure 15.

CASE NO.	1	2	3	4	5	6	7	8	9	10	11	12
1	78	72	90	88	090	87	82	74	80	46	083	60
2	90	88	78	78	077	50	72	48	66	46	087	60
3	84	82	78	80	080	57	72	60	60	60	073	67
4	60	74	74	86	087	90	76	62	90	58	090	83
5	82	74	84	80	090	80	88	74	82	42	097	93
6	50	50	54	44	083	83	60	66	72	64	087	73
7	74	84	86	72	090	83	74	46	64	66	097	73
8	78	80	86	90	077	60	84	80	80	58	100	67
9	82	64	94	86	093	70	78	84	86	66	083	80
10	70	60	80	64	083	67	88	54	70	56	093	60
11	88	80	90	86	090	93	90	66	82	54	087	93
12	74	80	90	72	093	73	78	58	80	62	083	67
13	84	88	82	84	100	77	72	50	72	62	083	93
14	64	66	86	86	090	90	74	78	70	70	067	90
15	56	62	72	78	087	67	70	66	68	74	097	73
16	84	68	92	84	077	77	80	84	72	48	093	83
17	82	92	98	82	080	67	86	68	64	54	083	67
18	90	82	84	78	067	80	78	78	66	70	083	87
19	74	70	82	62	090	60	79	76	72	48	087	50
20	88	88	90	74	087	73	90	72	68	60	077	80
21	56	68	62	70	067	60	74	52	82	58	073	50
22	60	58	94	90	083	70	78	70	68	52	090	67
23	72	52	44	54	083	70	62	58	46	54	087	50
24	60	52	60	72	077	57	74	64	72	54	053	27
25	60	78	90	76	073	77	76	68	82	76	061	60
26	70	78	86	80	100	67	78	68	82	64	093	97
27	72	76	80	86	090	77	78	68	66	54	087	67
28	76	68	86	74	093	67	48	64	78	70	090	53
29	74	60	88	62	087	73	72	74	72	56	100	80
30	86	68	80	80	093	73	88	58	78	64	100	87
31	56	58	80	70	083	87	68	64	82	66	077	73
32	82	62	88	72	083	80	80	70	76	56	100	83
33	86	72	72	74	097	83	76	70	92	70	087	83
34	52	52	94	72	087	80	82	64	72	60	073	70
35	68	78	78	66	080	53	84	62	82	64	083	47

Figure 16.

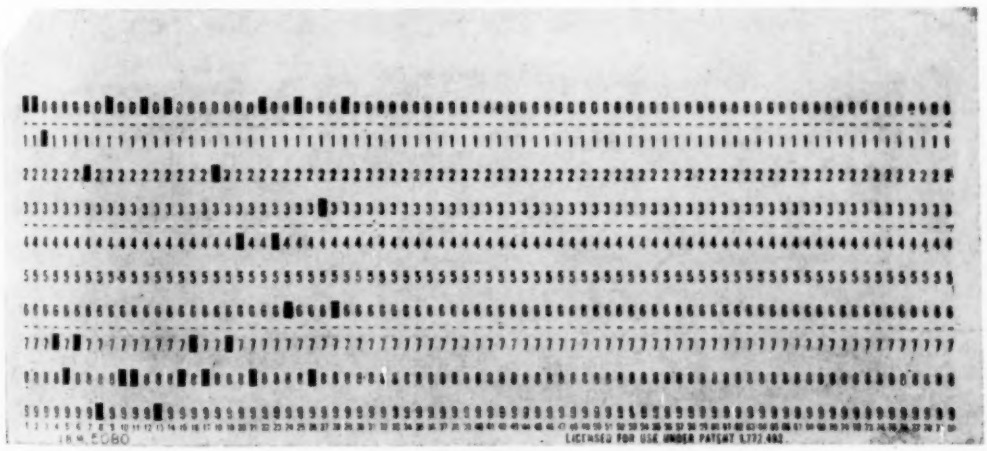
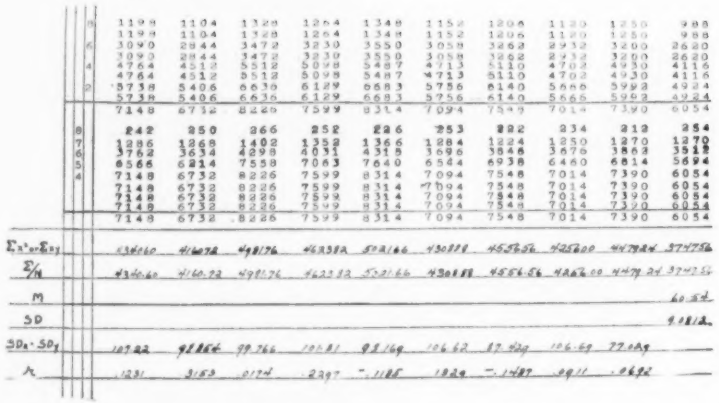


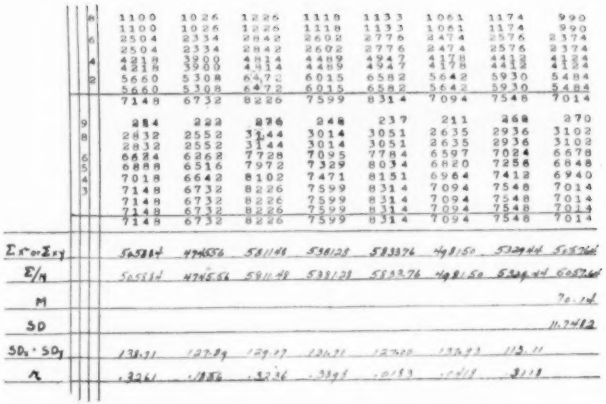
Figure 17.

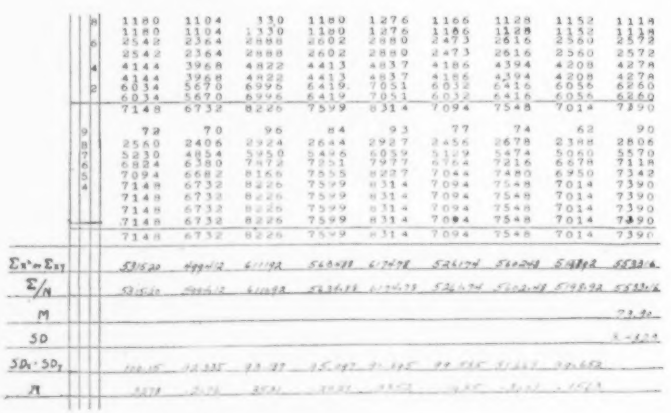
Residuals	Tech. Marks	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	7	2778	2616	3140	2956	3163	2717	2908	2640	2804	2302	3069	2536
		2778	2616	3140	2956	3163	2717	2908	2640	2804	2302	3069	2536
		2778	2616	3140	2956	3163	2717	2908	2640	2804	2302	3069	2536
		2778	2616	3140	2956	3163	2717	2908	2640	2804	2302	3069	2536
	5	4940	4600	5594	5194	5586	4800	5076	4724	4970	4102	5498	4633
		4940	4600	5594	5194	5586	4800	5076	4724	4970	4102	5498	4633
		4940	4600	5594	5194	5586	4800	5076	4724	4970	4102	5498	4633
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	6803
	9	868	832	1004	978	1043	886	926	930	900	768	1043	1106
	8	2718	2586	3108	2942	3130	2623	2790	2688	2774	2222	3178	3119
	7	3960	3638	4400	4158	4527	3856	4028	3752	3930	3144	4538	4314
	6	5354	5008	6056	5655	6140	5294	5528	5190	5454	4424	6185	5631
	5	6200	5838	7136	6567	7250	6200	6504	6064	6460	5188	7128	6307
	4	6558	6178	7552	6955	7653	6517	6888	6408	6824	5516	7458	6531
	3	6974	6560	8030	7403	8073	6920	7324	6824	7206	5886	7858	6742
	2	7034	6612	8090	7475	8150	6977	7398	6888	7278	5940	7911	6769
	1	7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	6801
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	6803
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	6803
Σx^2 or Σxy		494072	464104	565362	524666	572310	489939	519200	483312	507886	412934	561370	467265
Σ/N		4940.72	4641.04	5653.62	5246.66	5723.10	4899.38	5192.00	4833.12	5078.86	4129.34	5613.70	4672.65
M													68.83
SD													18.5620
$SD_x \cdot SD_y$		219.16	202.04	203.92	208.10	200.66	217.92	178.71	218.07	157.46	168.67	243.19	
r		.3556	.3032	.2810	.3703	.3298	.2906	.1516	.2361	.3267	.1641	.1135	

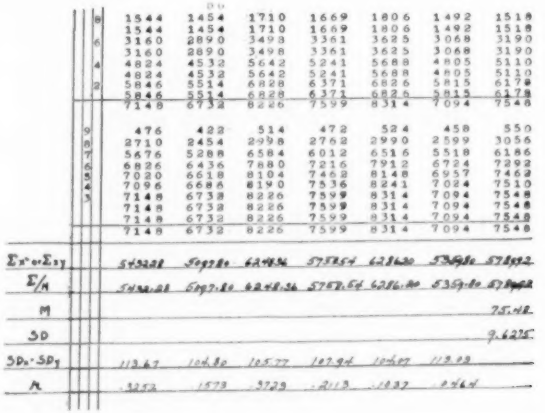
Figure 18.

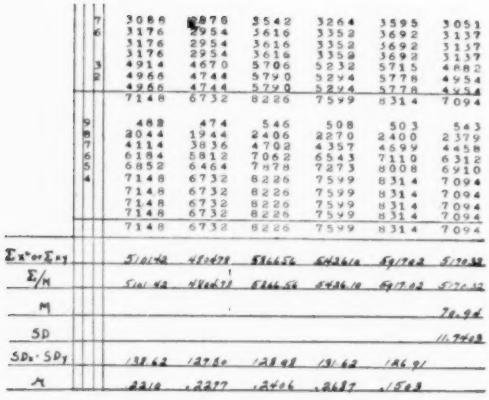
	7	1846	1744	2168	1972	2188	1825	1994	1790	1994	1564	2242	
		1846	1744	2168	1972	2188	1825	1994	1790	1994	1564	2242	
		1846	1744	2168	1972	2188	1825	1994	1790	1994	1564	2242	
		1846	1744	2168	1972	2188	1825	1994	1790	1994	1564	2242	
	3	4402	4226	5210	4786	5144	4406	4748	4372	4610	3794	5030	
		4402	4226	5210	4786	5144	4406	4748	4372	4610	3794	5030	
	1	4462	4304	5300	4862	5217	4483	4824	4440	4692	3870	5091	
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	
	9	2698	2450	2960	2738	3061	2571	2656	2498	2672	2074	3261	
	8	4752	4338	5368	4979	5447	4667	4842	4554	4794	3842	5662	
	7	5754	5312	6618	6105	6676	5735	6006	5554	5944	4762	6765	
	6	6196	5790	7172	6625	7229	6204	6468	6034	6420	5220	7192	
	5	6746	6334	7794	7181	7843	6750	7074	6580	6984	5712	7628	
	4	6874	6456	7938	7307	7986	6850	7234	6728	7124	5832	7711	
	3	6934	6526	8014	7387	8063	6917	7316	6790	7188	5900	7741	
	2	6988	6576	8074	7431	8150	6967	7384	6854	7236	5946	7761	
		6988	6576	8074	7431	8150	6967	7384	6854	7236	5946	7761	
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	
	1	160	156	152	168	164	127	164	160	154	108	200	
		7148	6732	8226	7599	8314	7094	7548	7014	7390	6054	7961	
Σx^2 or Σxy		575840	538912	659692	610962	666707	569575	602336	560804	593268	480854	650539	
Σ/N		5758.40	5389.12	6596.92	6109.62	6667.07	5695.75	6023.36	5608.04	5932.68	4808.54	6505.39	
M													79.61
SD													13.1011
$SD_x \cdot SD_y$		154.69	142.61	114.93	146.88	141.62	153.81	126.18	153.91	111.13	118.97		
r		.4420	.2021	.3346	.4089	.3410	.3135	.1141	.1572	.4454	−.0929		

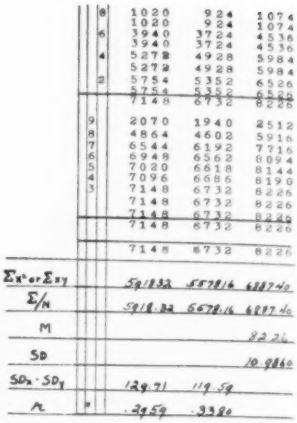


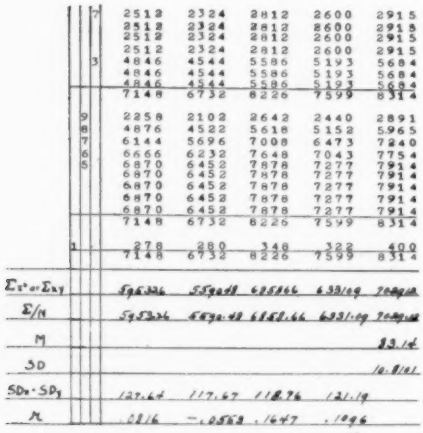


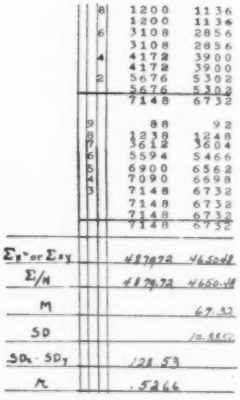


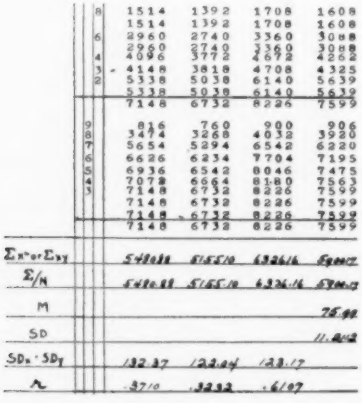


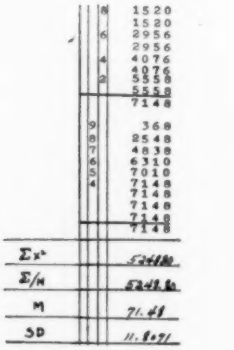












CORRELATION COEFFICIENTS AND OTHER CONSTANTS FOR COMBINED AND SEPARATED DISTRIBUTIONS

DOUGLAS E. SCATES
Duke University

There are numerous occasions when one wishes to obtain correlation coefficients and other series constants from several populations combined into one, or from a breakdown of a general population into two or more subpopulations. It is somewhat arduous to make new tabulations of the combined or separated populations. There are methods by which any number of tabulated distributions may be combined without additional tabulation or undue labor of calculation, and by which any tabulated distribution may be separated out from a compound table without further tabulation. The methods are equally applicable to calculations based on paired series which have not been tabulated, but for convenience of expression series will be spoken of as though tabulated.

A method of accomplishing this result was published some time ago by Dunlap,[1] and his method is called for when only means, standard deviations, and correlation coefficients are available. When however one has access to correlation work sheets the method herein presented[2] will probably be found more facile—especially when more than two subgroups are to be combined or separated. It is also highly flexible when certain special purposes apart from a simple combination are to be accomplished.

The method herein described is simply to sum, across the various groups to be combined, the respective totals which are normally required for the calculation of the desired constant in any single group, and then employ these sums in the formula in the usual way. For example, in the case of correlation for a single group, one commonly obtains six summations: N, $\Sigma x'y'$, $\Sigma x'^2$, $\Sigma x'$, $\Sigma y'^2$, and $\Sigma y'$—the primes indicating that an arbitrary origin has been used. To calculate the correlation for two or more groups combined (as though tabulated together on the same table), one needs only to combine the N's for the two or more groups, then combine the $\Sigma x'y'$ values for the two or more groups, and so on through the six summations. It is assumed only that the same working origins have been used for the different tables, and that the class intervals are the same; if they are not, corrections can be applied which will be described later. These combined summations are then used in the correlation formula just as though they had been derived directly from the compound table consisting of all the subgroups tabulated into one—for they will be the same values that one would get by actually combining the tabulations into one. If it is a compound mean or standard deviation that one wishes instead of a correlation, he combines only the summations which are needed, and substitutes them in the appropriate formula.

The procedure holds equally well for the correlation formula based on differences, $\Sigma(x' - y')^2$, instead of products, $\Sigma x'y'$: one sums the appropriate totals for that formula. In general, any form of correlation formula may be employed, so long as the working origins and class intervals are kept uniform, or the differences corrected for. Formulas which call directly for deviations from the actual mean of each subgroup do not keep the origins constant (unless all the means happen to be the same) and hence call for a correction appropriate to differing origins, as explained later.

The rationalization of the procedure may be set forth as follows: if several separate tables which are to be combined were actually combined by tabulating them into one, the frequencies in the subtables would in effect be combined cell by cell in the compound table. Assuming constant working origins and equal intervals, any one cell in the compound table will have the same deviation value as the corresponding cell in each of the separate tables. That is, the functional value of a single case

[1] Dunlap, Jack W. "Combinative Properties of Correlation Coefficients," *Journal of Experimental Education*, V (March, 1937), pp. 286–88.

[2] This procedure was worked out and described in the writer's doctoral dissertation: *Selective Admission and Selective Retention of College Students at the University of Chicago*, p. 107–30. Doctor's thesis. University of Chicago. 1926. 383 p.

in corresponding cells (in different tables) is the same cell by cell; and since all of the functions entering into the summations of a table are first degree with respect to frequency (the deviation values alone going to the second degree), the results for each cell or array are additive with respect to the frequency. In other words, a frequency of seven in a particular cell in the compound table will contribute *the same amount* to $\Sigma x'y'$ that separate frequencies of three and two and two, in the corresponding cell in three separate tables, will contribute when the $\Sigma x'y'$ values for the three separate tables are calculated and then added together; and the same is true of this frequency of seven with respect to the other five summations—$\Sigma x'^2$, $\Sigma x'$, etc.

Other ways of thinking about the combining may be simpler. In combining two tables one may tabulate them together on a distribution sheet, using different colored pencils for the two different populations. If now one calculates summations for the tallies of one of these colors, and then for the tallies of the second color, and adds these summations together, he will have the same totals as though he had counted both colors of tallies together. Yet the summations for each of the colors separately represent what he would have if he took summations from separate tabulations on separate distribution sheets. On the other hand, counting both colors together is what one naturally does when working from the compound table; so that the two processes are equal. Another way of seeing why summations of the functions come out the same by either process is to imagine that each of the tables to be combined is on a sheet of glass. The several subtables can therefore be laid one over the other: the frequency in any cell of the compound table would be the total which one would get by looking down through the several layers of glass and counting the tallies one can see in that particular cell in the several superposed tables. In similar fashion, the various marginal summations for a compound table would be seen to be sums of the corresponding figures on the several contributing tables. The calculation of r from the combinations of the several corresponding major summations is then exactly what one would do if all of the layers of glass were coalesced into a single table.

The relationship can be proved algebraically, though this would seem to add little. The proof amounts simply to expressing f_o, the frequency in the compound table, analytically as the sum of the corresponding frequencies of the contributing tables. Thus, for any one array (i.e., the same array, with respect to the working origin, in all of the tables),

$$f_o x' = (f_1 + f_2 + \ldots + f_k)x'$$

where the subscripts indicate the various tables. Therefore, for the entire correlation table,

$$\Sigma f_o x' = \Sigma(f_1 + f_2 + \ldots + f_k)x'$$

and the same thing is true for the other five summations ordinarily used. The additive character of these summation functions[3] with respect to frequency is thus seen.

Thus far the discussion has dealt only with combining. Subtracting a population out is simply the reverse. That is, if one has a correlation table which consists of more than one identifiable group, he can take out any group and calculate the correlation for the remaining group by subtracting the summations for the subgroup from the summations for the entire group: the remaining summations lead directly to the correlation for the remaining groups. The group which is thus taken out may of course itself be a compound of several subgroups. Successive groups may be taken out as desired; each group which is taken out as a unit however requires separate tabulation and calculation of its six moment summations. The economy of the present method rests in not having to retabulate one of the groups each time; that is, in taking groups out, one will seek to take out (and tabulate separately) the smaller group each time, obtaining the summations for the largest possible group by subtraction instead of by retabulating. Once the worker senses the freedom which is accorded him by virtue of the additive character of the moment summations he is dealing with, he can readily work out the best means of manipulating the groups.

Illustration With Data on College Admission

The data in Table I show how a number of combinations can be made economically if one plans his work with this in mind. The correlations represent the relationship between

[3] These six major summations for a correlation table are, strictly, *moments*, N being regarded as the zero moment (i.e., x is given an exponent of 0 and becomes 1) and Σxy being the product moment. It may be somewhat more expressive however to refer to them as the *moment summations*.

TABLE I

CORRELATIONS BETWEEN FOUR-YEAR HIGH SCHOOL AVERAGES AND AVERAGE GRADES IN FRESHMAN WORK AT THE UNIVERSITY OF CHICAGO

	Location of High School	r	N	$\Sigma x'y'$	$\Sigma x'^2$	$\Sigma y'^2$	$\Sigma x'$	$\Sigma y'$
	North Central States							
1.	Illinois*	.5622	553	5,689	10,580	9,409	352	289
2.	Indiana	.6210	261	3,070	4,545	5,536	65	336
3.	Ohio	.5019	83	774	1,457	1,695	9	93
4.	Michigan	.6565	98	1,052	1,554	1,670	26	—10
5.	Wisconsin	.5197	111	790	1,576	1,467	12	— 5
6.	Minnesota	.5529	23	313	680	400	—36	—50
7.	Iowa	.6122	194	2,139	3,607	3,728	—47	190
8.	All N. C. States*	.5750	1,323	13,827	23,999	23,905	381	843
9.	Atlantic States	.3366	110	724	2,497	1,875	—3	—57
10.	Southern States	.4756	106	907	1,614	2,227	34	133
11.	West Central States	.4996	168	1,470	3,689	2,591	—171	35
12.	Mountain and Pacific States	.6156	93	917	1,579	1,419	19	—13
13.	All States*	.5463	1,800	17,845	33,378	32,017	260	941
14.	Chicago City	.6120	1,707	17,063	30,081	28,979	1,719	—441
15.	Total Admissions	.5618	3,507	34,908	63,459	60,996	1,979	500

*Exclusive of high schools in Chicago.

four-year high school averages and averages made in the freshman year at the University of Chicago, for all high school students who entered the University during a five-year period.[4] The six summations shown for each row are values that one needs for the calculation of each r; and the adding of these values (vertically) gives the required summations for the correlation of a combined population. Separate correlation tables were initially prepared for each state. The correlation for the North Central states (line 8), for each of the groups of states (lines 9–12), for all states together (line 13), and for all admissions (line 15), were obtained without additional tabulation. Means and standard deviations for these groups were obtained in the same fashion. Thus, working with a total population of 3,500 cases (which had to be tabulated once, in various tables), the combinations shown were obtained with a saving of 7,100 case-tabulations over what would have been necessary if a new table had been made for each combination.

As an illustration of separating out populations, assume that one wished the correlation for all states together except Illinois. This value could be obtained by subtracting the summations for Illinois (line 1) from the summations for All States (line 13) and calculating the correlation from the remaining values—a few minutes' work, in lieu of tabulating 1,247 cases. Or, if one had not formed an All States total, he would add the summations for Illinois and for the City of Chicago (line 14) and subtract the total of these from the summations for Total Admissions (line 15).

[4] These data are from the writer's thesis previously cited. They have not heretofore been published, with the exception of the correlation for Chicago high schools which appeared in the following reference: Douglas E. Scates. "A Study of High-School and First-Year University Grades." *School Review* 32:182–92; March 1924. The data for out-of-Chicago high schools are for admissions during the five years 1919, 1920, 1921, 1922, and 1923; the data for the Chicago high schools are for admissions during the five years 1916, 1917, 1919, 1920, and 1921.

APPLICATIONS AND PRACTICAL CONSIDERATIONS

There are a great many ways in which one may subdivide and regroup correlation populations. The procedure described herein is adapted to combining or separating tables on the basis of a third variable (usually categorical), when the population is distributed on the same two variables in all of the tables. The division of the total group into tabulations according to categories or intervals of a third variable is equivalent, in principle, to partial correlation: but combining tables is not equivalent to multiple correlation. The latter assumes that various values are combined for each individual case before that case is entered into a single table; in the present paper we are discussing the compounding of entire tables, not individual cases. Subdivid-

ing a population on the basis of values in one or both of the two traits on which it is distributed, leading to "subcorrelations,"[5] is analogous to the present procedure only if one assumes that division on a particular third variable leads to contiguous but non-overlapping distributions on one or both of the two traits of the original correlation table. The procedure herein described has no relation to correlations between one or more component elements of a trait and the composite of all the elements—which normally utilizes a single population, and does not involve subgroups.[6]

The method will be of some help in successive regrouping of the data to accomplish either simple or joint subdivisions. That is, one may wish to break down a population on the basis of such traits as sex, age, location, etc. If these divisions are to be made independently, one will start with the original table for each breakdown; if they are to accomplish a joint subdivision, one will continue from the last preceding separation. In either case the procedure will save at least half of the retabulating for any dichotomous trait, and will save some work in other cases. If one foresees such analyses, he will probably proceed in the other direction, making separate tabulations for groups which are homogeneous with respect to sex, age, location, etc.; their successive combination into any number of compound tabulations is then simple and expeditious, involving no further tabulation. If one wants each correlation table fully tabulated, so that he can see the complete two-way distribution, then there is no saving to be found in the methods being discussed; but they will afford essential checks on one's tabulating and calculations, and may be employed solely for that purpose.

For those who are accustomed to utilize punch cards and Hollerith tabulating machines for the calculation of coefficients of correlation it may be pointed out that the combinations set forth herein may be obtained more quickly than cards can be sorted and tabulated on the machines to effect combinations of populations. Likewise when a large population has been correlated and is to be broken down into two very uneven groups, it may be found more expeditious to sort out the smaller group and obtain the correlation summations from that group on the machines, getting the summations for the larger subgroups by subtracting the small group summations from those of the total group. This is true even though one may be interested only in the larger of the subgroups.

Another situation exists when one has the means and standard deviations fully calculated, for each constituent population. For the present method this is no advantage, as completed constants are not utilized. The constants could be converted into summations, by employing N, but they would be summations around various true means and not around a common origin. They would therefore call for a correction (back to a common origin), or, if combined directly to form compound Σxy, Σx^2, Σy^2, and N, the final formula would involve two differences for each table (between its means and the means of the compound table). Starting with fully calculated constants is the situation for which Dunlap's formulas were derived, and if one begins with data in such form, he will be doing practically the same thing no matter which procedure he follows.

The theoretical or structural difference between Dunlap's method and the present one is that his procedure combines summations after they have been corrected to the subgroup means and then to the compound group means, whereas the present method combines them when taken around a common working origin, and corrects the summations to the compound group means in one step, in the final calculation. Both procedures yield the same results;[7] *the significant difference is purely a practical* one—*in what form are the data to start with?* If one has the necessary

[5] Dwyer, P. S. "The Use of Subcorrelations in Determining the Predictive Power of High-School Grades," *Journal of Educational Psychology*, XXVIII (December, 1937), pp. 673–80.

[6] Harris, J. Arthur. "The Correlation between a Component, and between the Sum of Two or More Components, and the Sum of the Remaining Components of a Variable," *Journal of the American Statistical Association*, XV (December, 1917), pp. 854–59.

Ghiselli, Edwin E., and Kuznets, George. "Short-Cut Methods for Calculating Raw and Corrected Correlations Between a Composite Variable and Its Components," *Journal of Educational Psychology*, XXVIII (March, 1937), pp. 237–40.

Conrad, Herbert S. "On the Calculation of the Correlation Between a Single Element of a Composite and the Remainder of the Composite," *Journal of Educational Psychology*, XXVI (November, 1935), pp. 611–15.

Edgerton, Harold A. "A Formula for Finding the Average Correlation of Any One Variable with the (n–1) Other Variables Without Solving Any of the Individual Correlations," *Journal of Educational Psychology*, XXVI (May, 1935), pp. 373–76.

[7] While Dunlap did not extend his formulas to cover subtraction of groups, this result is accomplished by subtracting (instead of adding) a set of parameters for the population in question, in both the numerator and the denominator of the formulas he presents.

summations, the present method obviates correcting them to the subgroup means and it corrects them to the compound group means in a single calculation instead of separately for each subpopulation. It does not entail the calculation of means or standard deviations for either the subgroups or for the compound group if these are not desired in themselves, for the method does not make use of them. If one has not planned his work so that the working origins and intervals are equal from table to table, and adjustments for these inequalities are necessary, the operations become more involved, and Dunlap's formulas may call for less calculation. It should be considered, however, that the present procedure permits readily of further combining and separating beyond the first operations, and after adjustments have once been made to bring working origins or class intervals in line these are no longer an obstacle. The formulas involving $Nr\sigma\sigma$, or Σxy, on the other hand, require the calculation of new means for the compound group, and the consequent changing of two differences in means for each table, for every subsequent combination or separation of tables.

Adjusting for Different Working Origins

If the same working origins were not used in calculating the deviations on the various tables, the summations for each population can be adjusted by the following procedure. If the origins actually used for a particular table are to be moved by the amount d_x or d_y in order to make them the same as the origins for other tables, the summations for the table will have the following corrections added to them (the sign of d varying inversely with the direction of the shift).

To $\Sigma x'y'$ add:* $\quad d_y\Sigma x' + d_x\Sigma y' + Nd_xd_y$

To $\Sigma x'^2$ add: $\quad 2d_x\Sigma x' + Nd_x^2$

To $\Sigma y'^2$ add: $\quad 2d_y\Sigma y' + Nd_y^2$

To $\Sigma x'$ add:** $\quad Nd_x$

To $\Sigma y'$ add:** $\quad Nd_y$

The sign of d must be preserved: if the new origin is smaller than the old, d is positive; if the shift is to a higher point on the scale, the amount of the change will have a negative value in the above expressions; but note that Ndd and Nd^2 will always be positive. The value of d must be expressed in the same units that the summations are: if they are in class intervals as units, then d must be similarly expressed before substituting in the corrections. The values and signs of d_x and d_y will of course be independently determined, according to the amount and direction of shift. If only one origin needs moving, to make it the same as the origins used in other tables, then the d for the other origin in that correlation table is zero, and two of the above five adjustments become zero, and the one for the product summation is reduced to a single term. The $\Sigma x'$ and $\Sigma y'$ which are called for in the first three adjustments are used just as they are obtained from the correlation table—that is, the values for $\Sigma x'$ and $\Sigma y'$ are *not* first corrected according to the two lower corrections, as that has been done already in the algebra. $\Sigma x'$ and $\Sigma y'$ adjusted by adding Nd to them are employed only for combining directly with other similar summations from other tables. The adjustments shown above

* If one is using a correlation formula employing a difference term in the numerator, as $\Sigma(x'-y')^2$, he may make the adjustment for a change in origin in any of three ways. It seems most expeditious to adjust for several terms of the numerator jointly. Since $\Sigma x'y' = (\frac{1}{2})[\Sigma x'^2 + \Sigma y'^2 - \Sigma(x'-y')^2]$, assuming the same origins and intervals, the adjustment given for $\Sigma x'y'$ may be applied directly to the three terms in the bracket as a unit, instead of adjusting them individually. It is permissible to do this for each separate table, before the values are combined with those of other tables, if one continues all the way through to treat this combination as a unit. Note that if this part of the numerator is not divided by two, twice the adjustment stated will be called for. The other term of the numerator, usually $\Sigma x'\Sigma y'/2N$, receives its attention in the final calculations, after $\Sigma x'$ and $\Sigma y'$ for the individual tables have been adjusted and combined.

A second form of adjustment for the term $\Sigma(x'-y')^2$ is to make an arithmetic combination of the values worked out for the first three expressions given in the list: since $\Sigma(x'-y')^2 = \Sigma x'^2 - 2\Sigma x'y' + \Sigma y'^2$, the desired adjustment may be obtained by adding together the adjustments calculated for $\Sigma x'^2$ and $\Sigma y'^2$ and subtracting from them two times the adjustment for $\Sigma x'y'$. A third method is to make this combination algebraically and calculate its value directly: $2(\Sigma x' - \Sigma y')\ (d_x - d_y) + N(d_x - d_y)^2$. If one is using a formula for r which contains a sum term, as $\Sigma(x'+y')^2$, similar plans may be employed, with appropriate changes in signs.

The symbolism Σd^2 is sometimes written in lieu of $\Sigma(x'-y')^2$. Such a d has no relation to the d used in this paper to represent the amount of voluntary shift in origins.

** The worker may note that, if d is taken equal to the distance from the working origin to the mean $(M - A)$, the value of the adjustment for $\Sigma x'y'$ becomes equal to the correction usually made in the product term (namely, $\Sigma x'\Sigma y'/N$) to correct it to the means of the group. The same thing is true of the adjustment made in the square terms of the denominator. Such equivalence shows the relationship between these formulas and the usual corrections, the present ones being more general; the two are equal only when d has the value stated, whence $d_x = \Sigma x'/N$ and $d_y = \Sigma y'/N$. One may note further that, when adjusted, the full value of $\Sigma x'y' - (\Sigma x'\Sigma y')/N$ remains the same as before the adjustment—and that the value of the denominator remains the same. This is as it should be—the correction must vary properly to offset the effect of changes in origin on the summations. The principal terms and the $\Sigma x'$ and $\Sigma y'$ must all be adjusted for each table however, in order for them to have the proper weight in the compound.

The value, Σd, as called for by the Cureton–Dunlap hand and machine correlation charts, and possibly others, has no relation to the d used in the present paper, but stands for $\Sigma(x'-y')$, or $\Sigma x' - \Sigma y'$. When it is called for, adjustments to it will automatically be obtained by adjustments to $\Sigma x'$ and $\Sigma y'$.

are essentially Charlier's expressions for checking by changing origins.

The summations may of course be adjusted for transferring their origin to any desired point, just so all the X variables in the several tables to be combined or separated have the same origin, and all the Y variables have the same origin. One may, if he desires, calculate the means for the compound table and use these as working origins. Little however is gained by so doing.

Adjusting for Different Class Intervals

If the several tables which are to be combined or separated have class intervals of different widths for the same variable, adjustments can be made in the moment summations by applying a ratio. The value of $\Sigma x'$ based on a class interval of 10 will (theoretically, at least) be just half as large as $\Sigma x'$ based on an interval of 5, and the value of $\Sigma x'^2$ will be one-fourth as large. To make changes in the summations for changes in the class intervals, one may proceed as follows:

If h_x and h_y represent the class intervals used in any table, and i_x and i_y represent the intervals which it is desired to change to, derive the ratios h_x/i_x and h_y/i_y. For simplicity of symbolism let these ratios be represented respectively by a and b. Then

Multiply $\Sigma x'y'$*	by ab
Multiply $\Sigma x'^2$	by a^2
Multiply $\Sigma y'^2$	by b^2
Multiply $\Sigma x'$	by a
Multiply $\Sigma y'$	by b

In other words, multiply the observed summation by the ratio of the width of its own class interval to the width of the desired interval, using the same degree of the ratio as the deviations have.

If one has to adjust the summations from a table for both class intervals and origins, it seems more straightforward to adjust for origins first, expressing d in terms of the old intervals. If however one makes the above adjustment for intervals first, the resulting $\Sigma x'$ and $\Sigma y'$ values will then be used in making the adjustments for changes in the origins as given earlier; also d will be expressed in terms of new intervals.

* The adjustment of $\Sigma(x'-y')^2$ may be accomplished by multiplying it by ab and then adding to it the following: $a^2\Sigma x'^2 + b^2\Sigma y'^2 - ab(\Sigma x'^2 + \Sigma y'^2)$. As in the case of adjusting for a change in origins, however, it is simpler to treat that part of the numerator, $(\frac{1}{2})[\Sigma x'^2 + \Sigma y'^2 - \Sigma(x'-y')^2]$, as a unit; the adjustment then becomes the same as for $\Sigma x'y'$, namely, multiplication by ab.

Eliminating Heterogeneity: Controlling Variation in Means of Subgroups

The procedure described in this paper is well adapted to accomplishing certain special purposes. Differences in means of subgroups may contribute positively or negatively (according to their position) to the correlation for the compound population—so that they may cause the resulting r to be higher or lower than any of the initial r's. If the combined population is regarded as homogeneous for one's purpose, this effect on the correlation is accepted as a part of the normal situation. If the combined population is not homogeneous for one's purpose, one may wish to remove from the compound correlation the effect of differences in means of the subgroups. This can readily be done while carrying through the procedures outlined. It can be done in one direction only, or for both of the traits at the same time.

To remove the effect of differences in means in one direction only (say the x means) of a number of subpopulations which are to be combined, one corrects the moment summations containing x for each individual table, back to the mean of that table as origin, before combining the summations of the several tables. For example, using the traditional correlation formula, one would apply to each individual table the full correction[8] for the numerator, $\Sigma x'\Sigma y'/N$, and also the correction for $\Sigma x'^2$ in the denominator, namely, $(\Sigma x')^2/N$. The summations for the several tables to be combined would then consist of: Σxy, Σx^2, $\Sigma y'^2$, $\Sigma y'$, and N. (Note the first two terms do not carry primes and are deviations from subgroup means.) After the adding together of these values for the several tables, r is calculated from the formula

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\ \sqrt{\Sigma y'^2 - (\Sigma y')^2/N}}$$

Differences in working origins for the X variable, from one table to another, need not be

[8] While this full correction does have the effect of correcting the numerator for differences in means in both directions, it is nevertheless not an excessive correction for one direction only. This follows because the correction for one direction only is the same as the correction for both directions. That is, when deviations are taken around the true mean of one trait, the entire correction becomes zero, regardless of whether the origin for the second trait is taken at the mean or elsewhere. This is obvious from the formula: if either $\Sigma x'$ or $\Sigma y'$ becomes 0, the entire correction, $\Sigma x'\Sigma y'/N$, becomes 0. This condition may be rationalized by noting that shifting the population orthogonally—along the direction of either axis (as when one origin is not at the mean)—does not have any effect upon the correlation, and hence does not affect the numerator.

corrected for, since the x' summations are corrected to the mean of each table as origin. Differences in class interval will be adjusted for as previously described, except that the $\Sigma x'$ term has become zero and needs no correction. It may be noted in passing that corrections could be made to some other point than the subgroup mean in the event that there was reason for doing so—such as when true subgroup means were known and the observed means were desired equated to them.

The above formula has the effect of correcting $\Sigma x'^2$ to the subgroup means and correcting $\Sigma y'^2$ to the mean of the compound group—the former being done before combining, and the latter after combining the moment summations from the several tables. (The signs of the various summations must be carefully preserved in the combining.) The results will be the correlation one would obtain if the several tables showed identical X means, but differed in their Y means. This is *not* the same thing as though all of the tables to be combined had their working origin in the X direction shifted to the actual X mean of the compound table—it is true that there would be no corrections called for in the numerator or in the Σx^2, as in the present case; but shifting a working origin has no relation to shifting actual means. Shifting the arbitrary origin has no effect upon the outcome; shifting the means so that they are equal produces a changed situation. If it is desired to equate both means of all the tables, then all three of the usual corrections will be made for each table and only Σxy, Σx^2, Σy^2, and N will remain to be summed. They will yield r directly without further corrections of any kind for origin, though adjustments for differences in size of class interval may be called for if different class intervals were used from table to table in obtaining the original subgroup summations.

One will make assumptions of these kinds, and resort to these procedures, when he desires to avoid introducing inappropriate heterogeneity into his compound population. It is a matter of somewhat difficult judgment to determine when this situation arises. One is always faced with the question, If the groups are not homogeneous when combined, why combine them at all? Why not calculate the separate correlations and stop? The combined population may iron out erratic fluctuations in a smaller group; but it may also cover up distinctive tendencies of distribution in the separate groups. The utilization of combining formulas therefore for groups which do not naturally belong together—especially when corrections are made for differences in one or both means—must be done after careful consideration. For example, one may have separate correlation tables for boys and girls. Boys and girls form a natural population when combined, and combining the tables is therefore legitimate. But to equate the means of the two groups before combining, by the methods just described, would raise a serious question. The result is a correlation that would obtain *if* the two sexes were equal on both means: but they are presumably not, and one therefore has the obligation of explaining why that correlation is as acceptable as the compound correlation based on observed (unequal) means.

The method of equating one or both means is available for partialling out the influence of an additional variable when that variable is chiefly responsible for changes in the means of one or both of the traits entering into the correlations. Thus, time may be held constant, when it is known to be the chief factor influencing the means of successive groups, and when the concommitance of variations in short intervals of time is of chief concern. Similarly age, or grade levels, of different groups of pupils may be held constant—if there is justification for so doing in a compound table. This type of partial correlation is not restricted to a quantitatively expressed third variable, nor to a linear distribution of means of subgroups. The individual distributions for specific values of the third variable give analytic, specific correlations; the compound correlation gives an average, similar to any partial correlation. It may be used where averages are what is wanted.

Controlling Variation in Dispersion Within Subgroups

The standard deviations of groups entering into a compound correlation do not raise or lower the resulting correlation directly, as means may do, but have the effect of weights. That is, they behave much like N does in this regard. They could not produce a compound correlation greater or smaller than any of the original correlations, as differences in means can; they simply weight more or less

the degree of correlation which is contributed by any one subpopulation. Being able to control the standard deviations may be of larger interest through enabling one to equate different scales of measurement that may have been used.

Equating, or removing the effect of variations in, the standard deviations of the different subgroups may be accomplished during the process of combining, and is similar to changing the value of intervals. The standard deviation in one trait only, or the standard deviations of both traits, may be controlled. First, we may assume that one wishes also to hold the subgroup means constant, and that the summations for each subgroup have been corrected to the means of that group. Then, to equate the standard deviations for one trait only, Σxy and Σx^2 for each subgroup may be divided respectively by σ_x and σ_x^2 for that group.[9] To control both standard deviations, Σxy is divided by the product of the 2 σ's, and Σx^2 and Σy^2 are divided respectively by σ_x^2 and σ_y^2—giving $\sqrt{N^2}$, or N, in the denominator. These divisions are made for each subgroup. Inasmuch as the subgroup standard deviations have all been made unity by these divisions, only two values need to be combined in order to obtain the compound correlation: N and Σxy. The correlation coefficient for the compound group is given by: $r = \Sigma xy/N$.

It is of course possible to equalize standard deviations by reducing all of them to some other value than unity; in this case the denominators will not reduce to N. It is also possible that one would not wish to equalize them, but to modify only selected ones. This offers no technical difficulties. All, or selected groups, may be multiplied by any desired ratio—so long as the ratio is squared for multiplying terms in the denominator which are under the radical.

In the foregoing the summations were all corrected to the subgroup means, which has the effect of equating the means. One may however wish to modify intragroup dispersion alone, without changing the means. Practically, the best procedure is to go through the foregoing steps for each group and then shift the summations concerned back to the common origin for all groups, by adding (for each subgroup) $\Sigma x'\Sigma y'/N$ back to the modified Σxy, in order to create a modified $\Sigma x'y'$ for the subgroup, and by adding $(\Sigma x')^2/N$ back to the modified Σx^2. All six of the summations for the various groups are then combined and r may be calculated. The reason for shifting the summations from the common origin to subgroup means, and then moving them back again to the common origin after multiplying or dividing, is that the deviations from an arbitrary origin contain a portion which must not be altered. While the $\Sigma x'^2$ and $\Sigma x'$ may both be changed by the same ratio for a single distribution, without error, they lose their proper weight when they are combined into a compound of summations.

There is one caution which should be observed. When one multiplies the moment summations for a given correlation he is doing more than changing its weight among a group of correlations. He is also multiplying the random errors in the original data and probably will not obtain the same numerator that he would from actual observations which had the same standard deviation as that to which he has mechanically changed his summations. This error may, under some conditions, be offset by utilizing Kelley's formula for the effect of range of talent on r, or Pearson's formula for curtailed distributions.[10] From one of these formulas, the appropriate subgroup r for the changed range may be obtained from the actual r for that table; when this is multiplied by $N\sigma\sigma$ (using the modified σ's which may have taken the value of 1), the desired modified Σxy is obtained for entering into the compound. If one is reducing standard deviations, the Σxy will be smaller than the one which would be obtained by directly reducing the original Σxy in proportion.

Alternative Procedures

For the purpose of simple combinations, the tabulating of separate tables for all desired combinations, either by hand or machine, and the use of Dunlap's combinative formulas, have already been commented on. It remains to note alternative procedures for holding means or standard deviations constant. Again, Dunlap's formulas are available; to hold the

[9] Subdividing Σx^2 by σ_x^2 yields N, the population of the subgroup. In reality therefore the division need not be accomplished except for purposes of checking; N may simply be substituted for Σx^2. For all of the subgroups combined, this gives the total population, which is used in place of Σx^2 in the final calculation of r. The formula becomes: $r = \Sigma xy/\sqrt{N\Sigma y^2}$, or $\Sigma xy/N\sigma_y$.

[10] Kelly, Truman Lee. *Statistical Method* (New York: Macmillan Co., 1923), pp. 221–30.

subgroup means constant, simply eliminate all of the terms containing delta in his formula 13; that is, the delta's all become equal to zero. The delta's may also be changed to some other value if desired. To equate or otherwise modify standard deviations, multiply the appropriate terms in the numerator by the desired ratio, and the appropriate terms in the denominator by the square of the ratio. Other special results as described in the present paper can be secured from his formula. The relationship of his formula to the procedure in the present paper should be clear when it is recognized that multiplying out such a term as $m\sigma_{x_m}\sigma_{y_m}r_{x_m y_m}$ gives Σxy for a subgroup; adding $m\delta_m\Delta_m$ shifts the origin to the compound group means. Similarly in the denominator, $m\sigma_{x_m}{}^2$ is Σx^2 for a subgroup, and adding $m\delta_m{}^2$ shifts the origin to the compound group mean. While the two procedures are interchangeable at various points, they start with different material, and make different demands. Dunlap's formula assumes completed calculations for all of the tables; the method of the present paper is based on raw summations, and in some cases tabulations may be entirely avoided.

May[11] presented a method of obtaining r when one starts with a compound table (or the essential parameters from it) and wishes to hold the means of the subgroups constant, without further tabulating. His method does not require linearity of the means, but requires their calculation. The method supplements that of the present paper, for it is adapted to use with (the results of) a compound correlation table and does not require any of the constituent tables, whereas the method herein described employs summations from the individual tables entering into a compound, and does not require a compound table. The choice between methods will therefore depend on where one is beginning in his work. May's method does not allow for modifying the individual standard deviations, and does not necessarily produce any tabulation.

McCloy[12] presented a method of obtaining a compound of analytical correlations which equates both the means and the standard deviations of the subgroups. It does not assume linearity, and could be adapted to equate only the means or only the standard deviations, as desired. The procedure builds up a compound correlation table through special tabulating so that means and standard deviations are held constant by the method of tabulating. It does not yield separate correlations for the various subgroups unless these are especially made. It will however yield eta.

Partial correlation, while usually thought of as analytical rather than combinative, is nevertheless an average—an average of specific or analytical correlations, with means equated; it is therefore comparable to the procedures described herein. It is definitely limited in two ways—through requiring linearity of means of subgroups when distributed on any pair of traits, and through requiring that the third variable be quantitative. It is true that arbitrary numbers may be used if the third variable is categorical, but these do not give satisfactory results unless they are so chosen as to lead to linear distributions of the means of subgroups or classes created by this third variable. Without this linearity the means of the classes are not held constant. The method does not lend itself readily to equating or otherwise modifying standard deviations. It is the only method of those discussed which is likely to prove practicable when more than four variables are employed. On the other hand, when the third variable is quantitative and linearity is fulfilled, the method is adapted to expressing the average relationship between the third variable and the first, for various specific values of the second trait. The methods already described do not yield such comparisons. For traits one and two, however, additional variables can be held constant by one method as well as the other.

All of these procedures will give the same results when the conditions which are assumed are fulfilled.

Summary

The correlation coefficients, standard deviations, and means may be obtained for combined or separated populations by adding together (or subtracting) the moment summations, for the several constituent groups, which are ordinarily obtained for individual correlations: $\Sigma x'y'$ or $\Sigma(x' - y')^2$, $\Sigma x'^2$, $\Sigma y'^2$, $\Sigma x'$,

[11] May, Mark A. "A Method for Correcting Coefficients of Correlation for Heterogeneity in the Data." *Journal of Educational Psychology*, XX (September, 1929, pp. 417–23.

[12] McCloy, C. H. "A Method of Computing Partial Correlation and Regression Equations with Variables Having Curvilinear Intercorrelations." *Journal of Educational Research*, XVI (November, 1927), pp. 285–95.

$\Sigma y'$, and N. After such combination (or subtraction) of the summations the results are substituted in the usual formula for calculating r from arbitrary working origins,—or in the appropriate formulas for the standard deviations or means. The procedure is not restricted to any particular correlation formula, nor to distributions which have been tabulated in two-way tables. It is somewhat more facile, when the necessary data are available, than are combining formulas which utilize directly r, σ, and M from each distribution.

If the working origins and class intervals are not the same for the tabulations of the several constituent populations, adjustments may be made for these differences.

The method is well adapted to accomplishing certain special purposes, such as eliminating variation in one or both means, and one or both standard deviations, of the several populations which are to be combined. These can be equated (held constant) by correcting appropriate moment summations for each population before these are combined. This has the effect of giving an average correlation for the several groups as though one or both of the means or standard deviations were equal. These are forms of partial correlation which are not dependent upon a quantitative third variable, nor upon assumptions of linearity of the distribution of means of the subgroups.